Universidad
Carlos III de Madrid

Modelo predictivo del coste de siniestros en automóviles aplicando Deep Learning

Ming-Da Liu Zhang

27 de Febrero de 2017

Agradecimientos

Quería expresar mi más sincero agradecimiento a todas aquellas personas que me han aportado un gran apoyo en este largo periodo de mi vida, que comenzó el lunes 5 de septiembre de 2011. La realización de este proyecto refleja los frutos de todo el esfuerzo dedicado durante estos seis años de carrera.

En primer lugar me gustaría agradecer a mis padres, Haoqing Liu Jin y Xina Zhang Zhou, y a mis hermanos, Chen-Da y Xiao-Da, por todo el ánimo y apoyo que me han dado estos años. Sin vosotros no habría sido posible.

Muchas gracias a mis tutores Irene Albarrán, Antonio Berlanga y Miguel Ángel Patricio, los que de principio a fin me han mostrado una completa disponibilidad, implicación y orientación que han permitido que este proyecto sea un éxito.

Agradezco también la Universidad Carlos III de Madrid y al Campus de Colmenarejo por haber hecho todo lo posible para ayudarme a crecer personal y profesionalmente.

Por último, me gustaría agradecer a a mis amigos más cercanos. Ellos han hecho más llevadera toda la carga de trabajo de estos años, y me han motivado día a día tanto en los buenos como en los malos momentos.

Tabla de contenidos

Índice de Ilustraciones

Índice de Tablas

Índice de Gráficas

Índice de Ecuaciones

1. Introducción

1.1. Contexto

Los automóviles son el medio más común en la sociedad actual para el transporte de personas o mercancías. Tal es la importancia del automóvil, que según el último Anuario de la Dirección General de Tráfico, en 2015 se han registrado más de 31 millones de automóviles en España. No obstante, el gran número de automóviles que circulan por nuestras carreteras generan una serie de riesgos que pueden afectarnos en cualquier momento. A pesar de los esfuerzos invertidos para el control de tráfico y la seguridad vial, todavía se producen un elevado número de accidentes, habiéndose registrado en el año 2015 un total de 97756, según datos de la Dirección General de Tráfico. Los daños que producen estos riesgos pueden ser personales o materiales, y la ocurrencia de estos mismos conlleva la realización de un siniestro que deberá soportar la persona afectada. El riesgo de accidente tan alto, junto a los grandes costes que conlleva exigen la existencia de un sistema compensatorio: el seguro de automóvil.

El conductor del vehículo causante de los daños causados en un accidente se debe hacer cargo de ellos, lo cual produce un daño patrimonial. Éste, para protegerse del riesgo patrimonial producido por el uso de los vehículos, contrata un seguro que le permite ceder el riesgo a un ente mayor, normalmente una compañía de seguros. El asegurado paga una cantidad de dinero llamada prima, y con ello consigue que la entidad aseguradora asuma las consecuencias económicamente desfavorables tras la ocurrencia de los riesgos asegurados, Debido a ello, es importante el cálculo y estimación de la misma.

Es de suponer que cada año estas empresas guardan en sus bases de datos información detallada de miles de peritajes de siniestros. Un buen manejo de esta información puede permitir mejorar tanto la toma de decisiones como la estrategia a determinar en la empresa aseguradora. A través de disciplinas como la Estadística o las Ciencias de la Computación, es posible el manejo de los datos y conseguir con ello información importante para ofrecer un servicio más adecuado a los clientes.

En el campo de las aseguradoras de automóviles, disponer de modelos fiables que puedan predecir las valoraciones en siniestros, peritaciones y análisis de casos de fraude es imprescindible. Según datos de la UNESPA (2015), el seguro de automóviles es donde se producen más de la mitad de reclamaciones fraudulentas (53%) de todos los casos de estafas a compañías aseguradoras.

Los nuevos modelos de seguros, al incluir todo tipo de modalidades y coberturas personalizadas para el cliente, junto con la creciente cantidad de información almacenada en las bases de datos, exigen el uso de nuevas técnicas y herramientas para manejar, analizar y predecir patrones sobre los clientes y los siniestros.

1.2 Objetivos

Este proyecto tendrá como objetivo aportar nuevas herramientas para predecir diferentes variables de interés de los siniestros que ocurran. En concreto se predecirá:

- El coste total que se ha tenido que pagar para la reparación completa del automóvil tras un siniestro.
- El coste total de las nuevas piezas necesarias para reparar el automóvil.
- Las horas de mano de obra requeridas para pintar el vehículo.
- Las horas requeridas para tratar la reparación de las chapas del automóvil.

Por otra parte, también se obtendrá información sobre qué datos influyen más a la hora de predecir estas variables de interés y qué datos son de menor utilidad.

Para cumplir con estos objetivos, una importante empresa de peritaje de seguros ha cedido una base de datos que contiene información detallada sobre 4.147.715 partes de accidente que corresponden a doce segmentos de vehículos (delimitado por marca, modelo, relación coste/calidad y tamaño). Estos datos contienen información sobre el coste total de las nuevas piezas, así como el coste total del siniestro y las horas de mano de obra de reparación de chapa y de pintura. A su vez, contienen información sobre el número de piezas en cada zona del automóvil, que se han tenido sustituir, reparar y/o pintar

Para ello, se utilizarán dos enfoques diferentes:

- Aprendizaje automático: a través de redes de neuronas con una estructura de aprendizaje profundo (Deep Learning) se usará el aprendizaje supervisado para predecir las variables de interés.
- Estadística: se usará el modelo de regresión lineal para la misma causa.

Estas herramientas a desarrollar tienen múltiples aplicaciones de interés para las empresas de seguros: se ayudará a la detección de fraude en el peritaje de siniestros de automóviles, dado que si la información de un parte de siniestro difiere exageradamente con los resultados predichos en los modelos, puede ser indicio de que haya algún intento de estafa; por otra parte, dará oportunidad de mejorar el cálculo de la prima a pagar en los seguros de automóvil, teniendo en cuenta la incertidumbre de las predicciones en los diferentes segmentos de automóviles.

Como que el Deep Learning ha obtenido bastante popularidad en los últimos años, también se realizará un breve estudio sobre cómo las redes neuronales se han extendido al aprendizaje profundo, y los algoritmos más utilizados de esta rama .

Por último, se realizará un estudio comparativo sobre los modelos predictivos desarrollados en cada disciplina (Estadística y Aprendizaje Automático), para determinar con ello las ventajas e inconvenientes de cada una y ver cómo se pueden complementar entre ellas para obtener mejores resultados.

1.3. Marco regulador

Los datos que se han cedido para la realización de este proyecto proceden de una entidad privada, y contiene información personal de los clientes y empresas. De acuerdo a la Ley Orgánica 15/1999, de 13 de diciembre de Protección de Datos de Carácter Personal, se han borrado todos los datos personales, de manera que la información restante queda completamente anónima frente a los que puedan realizar algún uso sobre ellas. Por otra parte, debido a la confidencialidad con la empresa que ha cedido estos datos a la universidad, no se incluirá en este estudio ninguna información sobre la empresa ni sobre los datos.

2. Situación actual

En este apartado se realizará una introducción al sector de los seguros, haciendo énfasis al seguro del automóvil. También se explicará el proceso de tarificación de seguros, junto al proceso de peritaje de siniestros de automóviles y el fraude en los seguros. Por último, se hablará sobre las disciplinas dedicadas al análisis de datos, haciendo enfoque a los algoritmos que se utilizarán en este proyecto.

2.1 Introducción al seguro de automóviles

2.1.1 Fundamentos del seguro

Siguiendo a Pérez Torres (2001), el seguro de automóviles se define como: "institución en que las personas que están expuestas a un riesgo agrupan sus recursos en un fondo común para hacer frente a las consecuencias económicas negativas que se producirán para aquellas en que el hecho constitutivo del riesgo ocurra realmente".

El seguro no evita la existencia del riesgo ni tampoco que la misma se produzca, pero a través de la unión de los recursos económicos de las personas expuestas al riesgo, se pueden compensar las pérdidas que sufren unos pocos afectados. De este modo, la aseguradora administra un fondo de dinero que acumula a través de las primas pagadas por los asegurados y paga con ello el coste de los siniestros.

Se distinguen entre diferentes clases de seguros (Pérez Torres, 2001):

- Seguros de personas: cubren riesgos que pueden afectar a la existencia, integridad personal o funcional o salud de una persona o un grupo de personas.
- Seguros de daños: tienen por objeto indemnizar al asegurado el perjuicio económico que sufre el interés que tenga dicho asegurado sobre un bien, en el caso de que ocurra el riesgo.
- Seguros patrimoniales: tienen por finalidad la cobertura de riesgos que, si acontecen, producirán una obligación para el asegurado, o le supondrán una pérdida que no afecta a un bien en concreto sino al conjunto de su patrimonio.
- Seguros multirriesgos: se incluyen pólizas que cubren diferentes riesgos a los que está sometida una actividad o bien asegurado.

2.1.2 El seguro de automóvil

El seguro de automóviles es uno de los seguros que más influye en la vida económica y social de los países desarrollados. La utilización de vehículos a motor es un elemento casi indispensable para el desarrollo de nuestra actividad diaria. No obstante, el gran número de automóviles que circulan por las carreteras generan una serie de riesgos que pueden afectarnos en cualquier momento.

Según el principio de responsabilidad civil definido en el Código Civil, en caso de un siniestro, el conductor del vehículo causante de los daños es el que se hará cargo de éstos. Esto quiere decir que la ocurrencia de los riesgos derivados del uso de vehículos a motor implica un daño patrimonial al responsable de los daños. El conductor, para protegerse del

riesgo patrimonial, contrata un seguro que le permite ceder el riesgo a una compañía de seguros.

El seguro de automóviles cubre los riesgos inherentes al uso y circulación de vehículos a motor (Guillén Estany et al., 2005). Este seguro está configurado sobre el vehículo a motor, y está limitado a la actividad de uso y circulación del mismo. Se consideran vehículos a motor: ciclomotores, motocicletas, turismos, vehículos comerciales (furgonetas), autocares, camiones, remolques, tractores, etc. En cuanto al uso y circulación de éstos, se entiende por hechos de circulación los derivados del riesgo creado por la conducción de los vehículos a motor en cualquier tipo de vías o terrenos (incluidos garajes y aparcamientos) aptos para la circulación o de uso común.

En España, la normativa relativa al seguro de automóviles es el Real Decreto Legislativo 8/2004 sobre uso y circulación de vehículos a motor. En la citada Ley, se reconoce:

- La importancia y necesidad del uso de vehículos a motor para el desarrollo individual y colectivo.
- El elevado riesgo que implica la circulación de automóviles sobre las personas físicas y los bienes materiales.
- La obligación de contratar un seguro de responsabilidad civil que cubra los daños ocasionados a terceros por la utilización de automóviles.

Esta Ley tiene como objetivo perseguir la indemnización inmediata de los daños y perjuicios sufridos por la víctima.

La responsabilidad civil mencionada anteriormente se define como la obligación que tiene una persona de reparar los daños y perjuicios ocasionados a un tercero (en su integridad física o en sus bienes) por causa de una acción u omisión.

El seguro de responsabilidad civil es un seguro patrimonial, dado que cubre el riesgo que corre el asegurado de que le reclamen responder con su patrimonio las personas dañadas por sus acciones u omisiones. Aun así, los seguros de automóvil suelen incluir coberturas para otro tipo de riesgos (salud, vida, daños al propio vehículo, etc.), por lo que suele tratarse de seguros multirriesgos.

Si bien, en un principio, el seguro de automóviles es un seguro de responsabilidad civil, existen algunos aspectos que lo separan del resto de seguros de este ramo (Guillén Estany et al., 2005).

- Principio de responsabilidad civil objetiva: admite la existencia de responsabilidad civil sin necesidad de que exista culpa o negligencia por parte de la persona causante del daño. Es decir, al contrario de la responsabilidad subjetiva que se basa la culpabilidad, la responsabilidad objetiva se basa en la causalidad. De esta forma, al no tener que determinar la culpabilidad de los implicados en un siniestro, se consigue de manera más rápida el resarcimiento inmediato de los daños y perjuicios sufridos por la víctima.

- Carácter obligatorio del seguro de automóviles: en la Unión Europea todo vehículo a motor debe estar protegido con un seguro que garantice a la víctima la indemnización de los daños causados por la circulación de éste.
- La determinación objetiva de los daños personales: la forma de evaluar los daños producidos difiere en los seguros de automóviles del seguro de responsabilidad civil general. En ambos, la determinación del perjuicio económico es subjetiva y concretatada por un juez. Sin embargo, para los daños ocasionados por accidentes de circulación, se reglas objetivas para valorar la cuantía a pagar. En España existe un Baremo de valoración para los daños personales derivados de los accidentes de circulación.

El objeto principal del seguro de automóviles es garantizar la responsabilidad civil del asegurado por el uso y circulación de un vehículo de motor. Existen dos tipos de responsabilidad civil: la obligatoria y la voluntaria. La responsabilidad obligatoria tiene como objetivo la cobertura a los daños personales o materiales causados a terceras personas. La responsabilidad voluntaria amplía los límites de la responsabilidad obligatoria y otras garantías.

2.1.2.1 Garantías aseguradas

A continuación, se describirán las posibles coberturas de un seguro de automóvil siguiendo a Guillén Estany et al. (2005). Hay que tener en cuenta que cada entidad aseguradora puede incluir en sus pólizas garantías o límites diferentes de las expuestas a continuación, de acuerdo al perfil del cliente, siempre y cuando cumplan con la legislación vigente. Así pues, las coberturas que pueden ser contratadas dentro de un seguro de automóviles son las siguientes:

Responsabilidad civil obligatoria:

Todo automóvil debe tener garantizada la cobertura del asegurado o conductor autorizado todas las posibles reclamaciones de terceras personas sobre las que se ha causado un daño personal o material. En España, la Ley sobre responsabilidad civil y seguro en la circulación de vehículos a motor regula el seguro obligatorio de automóviles, y está destinado a cubrir la responsabilidad civil del conductor del vehículo por las lesiones corporales o daños materiales que pueda causar a terceros por hechos relacionados con la circulación de su vehículo. En la Unión Europea y España, esta cobertura se limita de la siguiente manera:

- Por daños corporales, 350000 euros por víctima.
- Por daños materiales, 100000 euros por accidente.
- Los gastos de asistencia médica, farmacéutica y hospitalaria.
- Los gastos de entierro y funeral, en caso de muerte.

Sin embargo, la legislación excluye de esta cobertura todos los daños y perjuicios causados por el causante del siniestro, así como los daños sufridos por su vehículo o por las cosas transportadas en él.

Por otra parte, existen otros dos tipos de exclusiones en esta cobertura. El primero hace referencia a los daños corporales o materiales causados por una conducción bajo los efectos del alcohol o de otras sustancias estupefacientes. En este caso, el asegurador indemnizará al perjudicado, pero posteriormente podrá reclamar el coste del siniestro al asegurado.

La segunda exclusión hace referencia a los daños sufridos por un vehículo robado. En tal caso, la indemnización no corresponde al asegurador, que queda excluido de ésta, sino al Consorcio de Compensación de Seguros, que se encargará de indemnizar al perjudicado.

Responsabilidad civil voluntaria:

Las entidades aseguradoras ofrecen la posibilidad de complementar el seguro obligatorio de automóviles (SOA) a través del seguro voluntario de automóviles (SVA). Así pues, el SVA cubre un segundo tramo suplementario que sirve para garantizar la responsabilidad que pueda incurrir el conductor de un vehículo por los daños personales o materiales causados a terceros, para el caso en que la indemnización exceda las sumas garantizadas por el SOA. Aquí el asegurador garantiza el pago de estas indemnizaciones dentro de los límites marcados en el contrato.

Como en el seguro obligatorio, hay riesgos excluidos del SVA, como por ejemplo la responsabilidad civil por los daños materiales causados a los bienes transportados en el vehículo, la responsabilidad causada por los bienes transportados si el transporte del mismo no se adecúa al Código de circulación o multas y sanciones económicas. No obstante, estas coberturas se pueden incluir expresamente en el contrato.

Protección jurídica

Esta cobertura va muy ligada al seguro de responsabilidad civil, por lo que la mayoría de entidades aseguradoras las ofrecen conjuntamente. El seguro de protección jurídica cubre dos garantías:

- La defensa penal del asegurado o conductor responsable, con abogados y procuradores, el pago de los gastos judiciales y el depósito de fianzas.
- La reclamación de daños personales o materiales causados al asegurado por terceros.

Asistencia en viaje

La cobertura en asistencia de viaje incluyen como beneficiarios el tomador de la póliza, sus familiares, así como, el conductor y ocupantes del vehículo. Las garantías cubren gastos médicos, transporte, gastos de acompañantes, etc.

Este es uno de los complementos del seguro de automóvil que ha tenido mejor aceptación por los usuarios. Las garantías incluidas en esta cobertura son válidas y distintas para personas y vehículos.

Daños propios

Esta es una cobertura por daños propios para cubrir los daños materiales sobre el propio vehículo y puede comprender las siguientes garantías:

- Daños propios en el vehículo causadas por vuelco, caída o choque, hundimiento de terrenos, puentes o carreteras o falta o hecho malintencionado de terceros.
- Daños propios en el vehículo asegurado por incendio, explosión o rayo: para estas circunstancias quedan cubiertos los daños sobre el vehículo hasta los límites fijados en la póliza.
- Robo del vehículo asegurado y accesorios: robo del vehículo o partes de la misma, así como los daños ocasionados en el vehículo por robo o intento de robo.
- Rotura de cristales: gastos de reposición y colocación por rotura del parabrisas delantero, luneta posterior, cristales de las ventanas laterales y techo del vehículo asegurado.

Cuando hablamos del seguro a todo riesgo, normalmente se incluyen todas las garantías mencionadas.

Ocupantes

El seguro de ocupantes cubre las indemnizaciones, hasta el límite establecido en la póliza, que resultan de los accidentes de circulación en los que se producen lesiones corporales, incapacidad o muerte de las personas transportadas en el vehículo asegurado. El conductor del vehículo está igualmente cubierto por esta garantía, compensando así, la exclusión del conductor del vehículo causante en la cobertura de responsabilidad civil.

Otras garantías complementarias

Además de las garantías descritas hasta el momento, y sin perjuicio de las garantías acordadas según las condiciones particulares del asegurado, existen algunas coberturas complementarias. Por ejemplo, la cobertura en caso de retirada temporal del permiso de conducir, por la que el asegurado recibirá una indemnización mensual; la cobertura de equipajes, por la que el asegurador se compromete, hasta el límite fijado en la póliza, a indemnizar al asegurado de los daños o robo de los bienes transportados; o la cobertura de gastos de limpieza y acondicionamiento del vehículo en caso de daños producidos por el transporte de heridos, víctimas de un accidente de circulación.

Por otro lado, cuando la póliza del seguro de automóviles no incluye la garantía por daños propios en caso de accidente, pueden contratarse de forma complementaria, individual o conjuntamente, los riesgos de incendio, robo y rotura de cristales.

En la práctica comercial, las entidades aseguradoras ofrecen a sus clientes distintas modalidades para contratar su seguro de automóvil. En el siguiente esquema, se resumen los tres tipos de contrato más comunes:

- A terceros: incluye la responsabilidad civil obligatoria, responsabilidad civil voluntaria, protección jurídica, asistencia en viaje y ocupantes.

- A terceros con complementos: incluye todas las coberturas de un seguro a terceros e incluye otros complementos de daños propios (robo, incendio y/o rotura de cristales).
- A todo riesgo: incluye las coberturas de un seguro a terceros y todos los daños propios al vehículo.

2.2 Tarificación del seguro del automóvil

Una vez analizadas las garantías que puede albergar un seguro de automóviles, se analizará cómo determinan las entidades aseguradoras el precio o prima del seguro.

Guillén Estany et al. (2005), menciona que el precio final de un seguro deberá incluir por lo menos las siguientes variables: la prima de riesgo, destinada a garantizar el pago de los futuros siniestros; además, un recargo para hacer frente a los gastos de gestión, administración y comercialización; y, una serie de recargos externos, establecidos por la normativa vigente, que la entidad aseguradora recauda para otras instituciones. En este último apartado se incluyen los recargos para el Consorcio de Compensación de Seguros, el impuesto sobre primas (6%) y, el recargo para el Fondo Nacional de Garantía (3%).

La labor de determinar la prima de riesgo es bastante problemática en el ramo de automóviles, donde existen un gran número de factores que influyen en la probabilidad de ocurrencia del riesgo y un elevado número de pólizas, dando lugar a carteras de seguros muy heterogéneas. Por ello, los actuarios utilizan numerosas herramientas con el fin de determinar y valorar el riesgo que cada póliza aporta a la cartera de seguros de la entidad.

Existen dos fases en la tarificación de pólizas: la tarificación *a priori* y la tarificación *a posteriori*. A continuación se explicará cada tipo de tarificación siguiendo a Guillén Estany et al. (2005).

2.2.1 Tarificación a *priori*: segmentación

Este método consiste en segmentar los riesgos en clases homogéneas de forma que todos los asegurados que pertenecen a una clase paguen la misma prima. Con ello se intenta solventar el problema de la heterogeneidad en la cartera de seguros.

Antes de firmar el contrato, es preciso que la entidad aseguradora pueda medir el riesgo asegurado y el daño que puede provocar. En este punto, el asegurador necesita predecir el comportamiento del riesgo a partir de la experiencia pasada del conductor. Como no todos los riesgos del mismo ramo tienen la misma probabilidad de ocurrencia o provocan la misma intensidad, las entidades aseguradoras deben determinar los factores o características que afectan a cada riesgo y clasificarlos o de acuerdo con estos factores de riesgo. Las variables de clasificación más comunes son las siguientes:

- Los relativos al vehículo: por ejemplo, la categoría del vehículo (vehículos hasta 3500 kg, de más de 3500kg, motocicletas, etc.), la marca y modelo, la potencia, el color de automóvil, etc.

- Los relativos al conductor: normalmente, la edad y sexo del conductor, la antigüedad de su permiso de conducción, el estado civil y número de hijos, entre otros.
- Los relativos a la circulación para: para tener en cuenta el uso del vehículo (articular, profesional, de reparto, de alquiler, etc.) y la zona de circulación de éste, para tener en cuenta factores como las infraestructuras, pluviosidad, geografía, densidad de tráfico, o si se trata de zona rural o urbana.

2.2.2 Tarificación a *posteriori*: bonus-malus

Existen importantes factores de riesgo que no son tenidos en cuenta en la tarifa a *priori*, puesto que son muy difíciles de cuantificar, como son: los reflejos y agresividad del conductor, el conocimiento del código de circulación, etcétera. En consecuencia, las clases de tarifas siguen siendo bastante heterogéneas. Por ello surge la idea de tener en cuenta estas diferencias a *posteriori*, ajustando la prima individual en función de la experiencia individual de siniestralidad de cada asegurado.

Este sistema de tarificación consiste en conceder bonificaciones o reducciones en la prima a los asegurados según los siniestros que sufran. Es decir, se modifica la póliza inicial según se vaya recogiendo más información del asegurado una vez iniciada la relación mediante el seguro.

El sistema bonus-malus es beneficioso tanto para el asegurador, garantizándole unas primas suficientes de acuerdo con el riesgo asumido, como para el asegurado, que de esta manera paga el precio justo por el riesgo que decide asegurar.

No obstante, para que este sistema funcione correctamente es imprescindible que las diferentes entidades aseguradoras tengan un acuerdo en el que pongan en común todas las estadísticas de siniestralidad de cada asegurado. En caso contrario, el asegurado que se encuentre en la situación en que le vayan a subir la prima, podría abandonar su compañía aseguradora por otra que, desconociendo sus datos, no le aplicaría el recargo de precio. En tal escenario, las compañías sólo concederían tarifas rebajadas de precio, lo cual podrían suponer una quiebra de las aseguradoras debido a la insuficiencia de las primas.

Por esta razón, entre las entidades que operan en el sector, se ha creado un fichero histórico de siniestralidad de conductores (SINCO), que incluye el historial de siniestralidad de los últimos cinco años de cada póliza.

2.3 Siniestros de automóviles

Un siniestro es el acaecimiento del hecho previsto en la póliza que desencadena el cumplimiento de las obligaciones del asegurador (Guillén Estany et al., 2005). En el seguro de automóviles, las obligaciones de la entidad aseguradora surgen de distintas causas, según el tipo de garantía cubierta. Por ejemplo, pueden consistir en el pago de una indemnización por responsabilidad civil, daños propios y robo, reparación de los daños materiales causados o en la prestación de servicios.

2.3.1 Gestión del siniestro

El tratamiento de los siniestros realizado por el asegurador es un proceso largo y complicado, en el cual se pretende aclarar los hechos, causas y circunstancias del siniestro, así como evaluar las consecuencias económicas del mismo. A continuación, se describen las cuatro fases que comprende un siniestro (Guillén Estany et al., 2005):

1. Declaración del siniestro: se trata de la primera actividad después de haber ocurrido un siniestro, en el que el asegurado comunica de forma verbal o escrita el siniestro a la aseguradora. La aseguradora puede pedir al asegurado toda la información y documentación que considere conveniente. Si el asegurado se niega a proveer dicha información, puede perder el derecho a la indemnización.
2. Tramitación del siniestro: tras la declaración de siniestro, la entidad abre el expediente correspondiente. Aquí se incluye: el número de siniestro, la persona que ha recibido la declaración y la que tramita el siniestro, el número de la póliza afectada y los datos del asegurado, así como las circunstancias y consecuencias del siniestro. Acto seguido, la compañía debe hacer una primera valoración del coste del siniestro. Esta primera valoración suele coincidir con el coste medio de los siniestros de la misma categoría en que se encuentre. Por otro lado, la aseguradora comprueba si es ella la que tiene que hacer frente a la indemnización, verificando que el asegurado haya pagado la prima y que el siniestro cumpla con todos los requisitos que figuran en el contrato. Si el siniestro no está cubierto, la entidad rechaza el pago de la indemnización y finaliza el proceso de tramitación, cerrando el expediente.
3. Peritación del siniestro: si la entidad lo cree conveniente, puede decidir encomendar la valoración del coste del siniestro y la investigación de éste a un profesional. Este profesional, el perito, tras estudiar el siniestro, realiza un informe pericial. De este informe depende que el asegurador decida pagar los daños y en qué cuantía. No obstante, si el asegurado no está de acuerdo con lo establecido en el informe pericial puede impugnar por vía judicial.
4. Liquidación del siniestro: una vez se ha valorado el coste del siniestro e investigadas las características de éste, la entidad puede pagar la indemnización correspondiente o negarse a pagarla. Si se niega a pagar, debe notificar por escrito al asegurado y éste podrá reclamar el pago por vía judicial. se requiere la notificación escrita al asegurado y éste podrá reclamar el pago de la indemnización por vía judicial.

2.4 Fraude en el seguro del automóvil

La lucha contra el fraude se ha convertido en uno de los principales objetivos de las entidades aseguradoras, fundamentalmente, de las que operan en el ramo del automóvil (UNESPA, 2015).

Guillén Estany define el fraude como "el intento de ocultar circunstancias o de distorsionar la realidad para obtener un beneficio más allá de la justa compensación″ (2005).

En los contratos de seguros, el fraude tiene una componente de aleatoriedad que no se da en otros contextos. Sólo se puede obtener un beneficio ilícito en el momento de cobrar una

indemnización y, ésta, sólo se podrá conseguir si se produce un siniestro. Pero el siniestro no es siempre previsible y, por ello, puede ocurrir que el asegurado decida defraudar cuando se produce el accidente siendo éste real, o bien, puede fingir la ocurrencia de un accidente que nunca se ha producido.

Tradicionalmente, el fraude se ha considerado como un suceso inevitable, por lo que la solución que usan las aseguradoras es la de aumentar el coste de sus primas, repartiendo el coste al resto de los clientes y compensando así las pérdidas ocasionadas. Por ello se están diseñando herramientas que ayudan a peritos y tramitadores a mejorar, y en medida de lo posible, a sistematizar la investigación del fraude.

2.4.1 Tipología del fraude

Todos los casos de defraudaciones o intentos de fraude tienen en común el falseamiento u ocultación de datos o circunstancias en la declaración o tramitación de un accidente, con la finalidad de que el asegurado o un tercero obtenga una indemnización que, de otro modo, no le correspondería. A continuación se exponen los casos más frecuentes de fraudes (Iturgoyen, sf):

- Falsa declaración para favorecer a un tercero: es el caso más usual de fraude. Esta circunstancia está presente en el 30% de los casos. Consiste en el falseamiento de la declaración de siniestro que realiza un asegurado para <<hacer un favor>> declarándose <<culpable>>.
- Falsa declaración del asegurado para eludir casos excluidos en la póliza: este caso es también muy frecuente y está presente en el 27% de los fraudes detectados. Comprende todos los casos en los que al declarar el accidente se ocultan hechos o circunstancias cuyas consecuencias quedan excluidas de la cobertura del seguro.
- Falsa declaración para obtener un beneficio el propio asegurado: en esta modalidad es el propio asegurado el que pretende beneficiarse directamente y se detecta en el 11% de los fraudes descubiertos.
- Contratación de la póliza después de ocurrido el accidente: se detecta en el 6% de los fraudes e implica el engaño y/o complicidad de algún empleado u agente en la entidad aseguradora.
- Otras modalidades como ocultación de alcoholemia (3%), falso conductor habitual (2%), fraude del taller (2%) y versiones para cobrar ambos implicados (1%).

En ocasiones es muy complicado determinar las causas de un siniestro y los indicios de que haya fraude, por ello, muchas compañías aseguradoras se fijan en determinados aspectos que pueden llamar la atención y dar motivos de investigación:

- A través de la peritación: la utilización de esta vía o en combinación con otras, permite la detección del 61% de los fraudes. Al efectuar la peritación se comprueba que: los daños no se corresponden con la mecánica del accidente; la configuración de los daños (alturas, trayectorias, etc.) no coinciden; los restos de pintura no son los indicados; la antigüedad de los daños (óxidos, suciedad, etc.) prueba que éstos

se ocasionaron con anterioridad a la fecha declarada; o se aprecian restos extraños (hierbas, cemento, tierra, etc.) que no tienen que ver con el relato del accidente.

- Por el relato del accidente: es la segunda causa más importante como circunstancia sospechosa, constituye en combinación a otras técnicas el 52% de casos detectados.
- Daños muy elevados en el vehículo contrario: este caso se detecta en el 25% de los fraudes descubiertos.
- Por nerviosismo o contradicciones en la declaración del accidente: con esto se detecta el 9% de los fraudes.
- Otras: con menores porcentajes, aparecen otras circunstancias tales como fecha y hora del accidente (6%), emisión de la póliza próxima al accidente (5), indicaciones del taller (4%), etc.

2.5 Minería de datos

El avance en la recogida masiva de datos y en el almacenamiento de las mismas ha dado lugar a bases de datos gigantescos. Debido a ello, actualmente ha surgido la necesidad de usar esta inmensa cantidad datos y extraer de ello toda información que pueda ser valiosa. La disciplina que se encarga de esta tarea es la minería de datos.

Hand, Mannila y Smyth definen esta disciplina como: "el análisis de grandes volúmenes de datos con el fin de encontrar patrones no triviales y resumir los datos de manera comprensiva para que éste sea entendible y útil." (2001).

La minería de datos se compone de cuatro fases (Hand et al., 2001):

- Selección de los datos: partiendo de unos datos en crudo disponibles, se identificarán las variables objetivo (aquellas que se quiere predecir, calcular o inferir) y las variables de independientes (las que servirán para realizar los cálculos necesarios).
- Preprocesamiento de datos: se determinan qué datos son útiles y se eliminarán los datos irrelevantes, erróneos y atípicos. Posteriormente, se realiza un nuevo tratamiento de los datos para dejarlos preparados y adecuar para su uso en los algoritmos a utilizar. En este tratamiento realizan operaciones tales como la normalización, aleatorización, reducción de dimensiones, separación de datos en subconjuntos (conjunto de entrenamiento y conjunto de test), etc.
- Extracción del conocimiento: se realiza el proceso de minería de datos y se identifican patrones con el objetivo de realizar una clasificación, regresión, agrupamiento, asociación, etc.
- Evaluación e interpretación: en esta fase se analizan los resultados obtenidos y se verifican que éstos son coherentes, elaborando así un modelo final. Estas cuatro fases se repiten hasta obtener los resultados deseados y obtener un modelo lo suficientemente preciso.

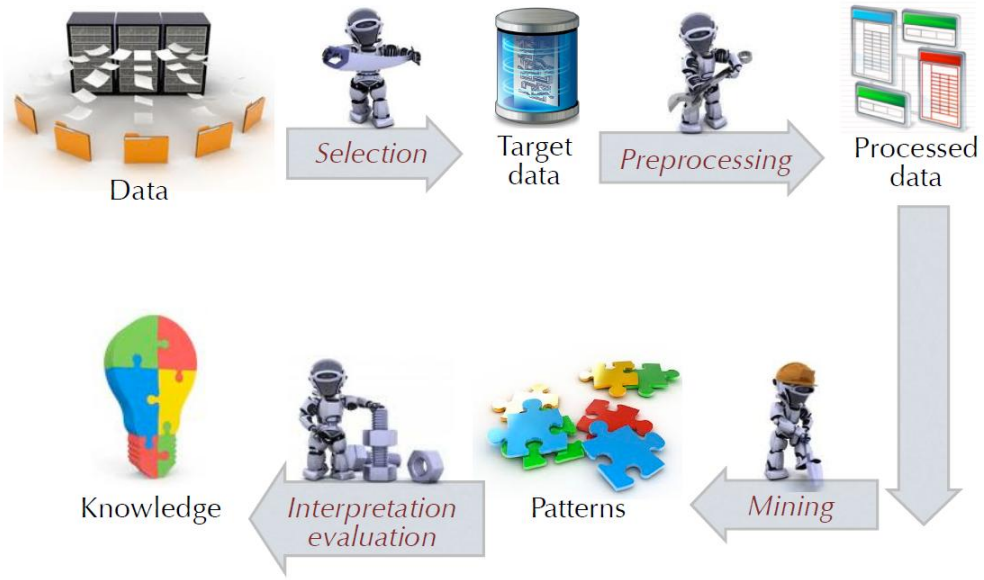


Ilustración 1: Proceso de la Minería de datos

Por último, dentro de la minería de datos se diferencian dos objetivos distintos (Gorunescu, 2011):

- Objetivos predictivos: en este tipo de problemas, se dispone de un conjunto de variables conocidas, y se busca la relación que tengan éstas con otras variables desconocidas con el fin de predecir su valor. Dentro de esta categoría nos podemos encontrar con problemas de clasificación, regresión, predicción o análisis de series temporales.
- Objetivos descriptivos: se buscan patrones comunes para agrupar los datos en diferentes conjuntos y comprender las características de cada uno. Aquí nos podemos encontrar con problemas de agrupamiento, reglas de asociación y descubrimiento de patrones de secuencia.

2.6 Aprendizaje automático

El aprendizaje automático o *Machine Learning* es un campo dentro de la inteligencia artificial que busca estudiar y desarrollar algoritmos que puedan aprender y realizar predicciones sobre los datos. "Se trata del campo de estudio que busca dar a los ordenadores la capacidad de aprender sin tener que realizar explícitamente un programa para ello" (Samuel, 1959) .

Con el aprendizaje automático no sólo se intenta construir algoritmos para el problema a resolver, sino que también busca dar respuesta a situaciones desconocidas previamente y mejorar la eficiencia de las tareas propuestas.

Esta disciplina se usa principalmente en tareas en el que programar algoritmos de manera manual es extremadamente difícil. Algunos ejemplos en los que destaca el aprendizaje automático son problemas como el filtro de spam, reconocimiento de imágenes, detección de fraude, segmentación de clientes, etc.

Los tipos de aprendizaje automático más importantes son el aprendizaje supervisado y el aprendizaje no supervisado. Este proyecto se usará en el aprendizaje supervisado.

2.6.1 Aprendizaje supervisado

Es la tarea del aprendizaje automático que busca inferir un modelo a partir de un conjunto de datos denominados datos de entrenamiento en el que se conoce su valor o resultado. Los datos de entrenamiento consisten en un conjunto de ejemplos, en el que cada ejemplo consiste un par de valores compuesto por unos datos de entrada y una salida conocida. Un algoritmo de aprendizaje supervisado analiza el conjunto de datos de entrenamiento y produce una función que se usa para predecir el valor de nuevas instancias. La salida de la función será un valor numérico (problema de regresión) o una clase (problema de clasificación). Se busca un escenario en el que el algoritmo tenga la capacidad de determinar correctamente qué valor tienen instancias que no se han visto previamente, es decir, que tenga capacidad de generalizar. Este tipo de aprendizaje automático se corresponde con el objetivo predictivo de minería de datos de la que se ha hablado en el apartado 2.5.

Para ver en qué medida el algoritmo ha conseguido aprender a partir de los ejemplos, se dispone de un segundo conjunto de datos denominado conjunto de test con el mismo tipo de datos. El algoritmo utilizará los datos de entrada para realizar sus predicciones, y comparará los resultados con la salida real. De esta manera es posible comprobar en qué medida el modelo está obteniendo los resultados deseados.

Un problema típico a las que se enfrentan los algoritmos de aprendizaje automático es la del sobreajuste de los datos (Dietterich, 1995). Esto ocurre cuando el algoritmo predice muy bien en el conjunto de datos de entrenamiento, pero no es capaz de generalizar el aprendizaje a nuevos casos. Como consecuencia, se obtienen resultados pobres en el conjunto de test o de validación, tal y como se muestra en la ilustración 2.

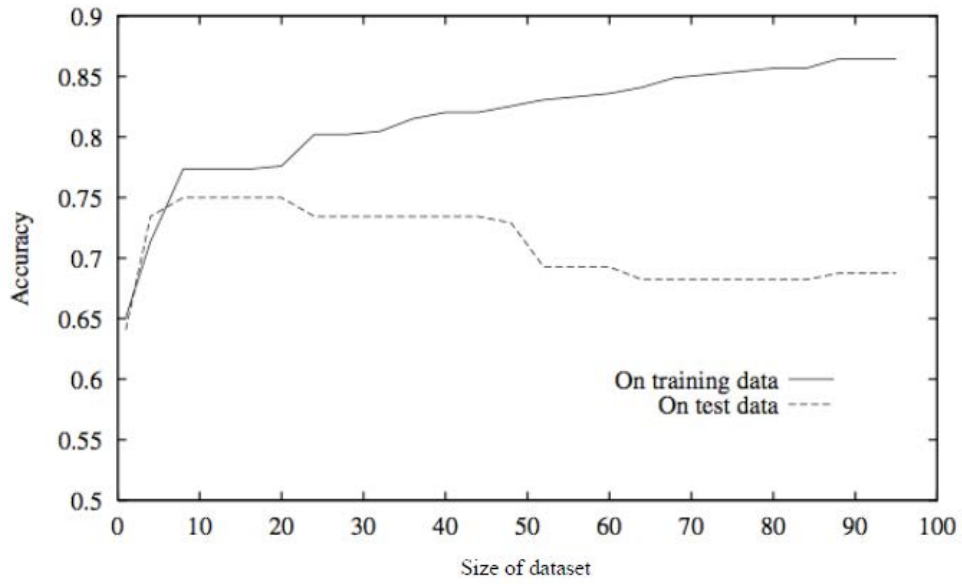


Ilustración 2: Ejemplo de sobreajuste de datos

2.6.2 Redes de neuronas artificiales

Las redes de neuronas son una herramienta de aprendizaje automático que se basa en el funcionamiento del cerebro humano: utiliza un conjunto de unidades neuronales, en donde cada neurona está conectada con muchas otras neuronas, y a través de sus enlaces, envían una señal con información sobre su estado, lo que provoca la activación de las neuronas adyacentes. Cada unidad neuronal es un elemento simple que realiza una operación de transformación en la que combina todos sus valores recibidos y produce con ello una salida.

Hecht Nielsen define las redes de neuronas de la siguiente manera:

"...un sistema de computación hecho por un gran número de elementos simples, elementos de proceso muy interconectados, los cuales procesan información por medio de su estado dinámico como respuesta a entradas externas" (1995).

Gracias a la interconexión de múltiples elementos de proceso (neuronas) y a los cambios adaptativos de sus parámetros, se consiguen realizar operaciones muy complejas. Este tipo de sistemas se basan en el auto-aprendizaje y destacan en problemas en donde es difícil diseñar una solución mediante el uso de los métodos tradicionales de programación.

Las redes de neuronas ofrecen múltiples ventajas frente a los enfoques tradicionales de programación (Haykin, 2005):

- Capacidad de resolver problemas complejos y no lineales, típicos de problemas provenientes de la naturaleza como el reconocimiento de voz.
- Adaptabilidad: las redes de neuronas tienen la capacidad de adaptarse dependiendo del entorno en que se encuentre. Aprenden a realizar sus tareas en base a una experiencia inicial.
- Auto organización: para calcular las salidas o resultados la red neuronal realiza sus operaciones internamente, lo cual ahorra al usuario el trabajo de diseñar relaciones y reglas explícitas para la solución.
- Tolerancia a fallos: tiene la capacidad de ser robusto, en el sentido que si una parte de la red está fallando, el rendimiento de la misma va disminuyendo poco a poco, y no de golpe.

2.6.2.1 Fundamento de las redes de neuronas

En este apartado se describirán los elementos más importantes que forman las redes de neuronas siguiendo a Haykin (1995).

Una neurona es la unidad principal de procesamiento de una red neuronal. La ilustración 3 muestra el modelo básico de una red neuronal. Aquí se identifican 3 elementos básicos:

- Un conjunto de enlaces que conectan las diferentes neuronas entre sí. Cada enlace está caracterizado por tener un peso, que indica en qué manera la neurona afecta a a su adyacente. Es decir, la neurona k recibe señales de entrada con valores x_j a

través del enlace j, en donde modifica la intensidad de la señal al multiplicarlo por su peso w_{kj} .

- Una función sumatoria que se encarga de agregar todas las señales de entrada.
- Una función de activación con el objetivo de limitar el valor de salida de una neurona a un rango determinado. Normalmente interesa que el rango de la salida de una neurona se encuentre en el intervalo [0,1] o [-1,1]

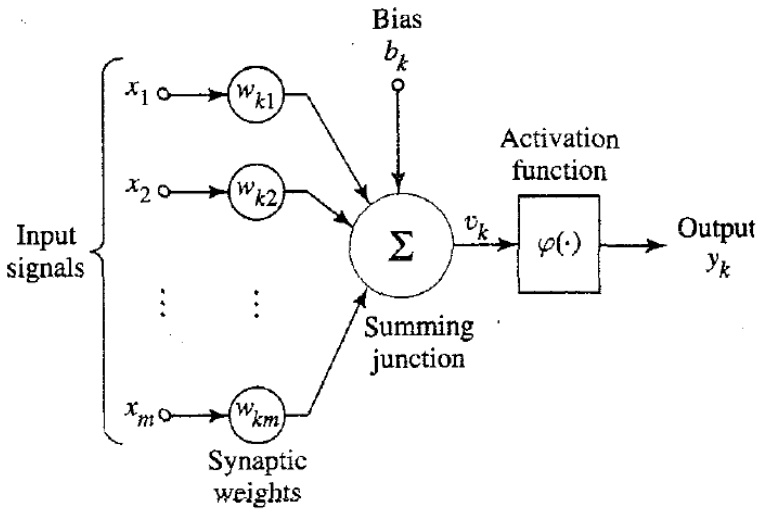


Ilustración 3: Modelo de una red neuronal

Matemáticamente se describe a una neurona k con el siguiente par de ecuaciones:

$$u_k = \sum_{j=1}^{m} w_{kj}\, x_j$$

Ecuación 1: Función de agregación

$$y_k = \varphi(u_k + b_k)$$

Ecuación 2: Salida de una neurona

En donde $x_1 x_2, \dots, x_m$ son los valores de entrada; $w_1 w_2, \dots, w_m$ son los pesos de la neurona k; u_k es la salida agregada de los valores de entrada; b_k es el sesgo; $\varphi(.)$ es la función de activación; y y_k es el valor de salida de la neurona. El uso de b_k se utiliza para ajustar la transformación de la salida u_k, tal y como se muestra en la Ilustración 3:

$$v_k = u_k + b_k$$

Ecuación 3: Valor de entrada de la función de activación

Tipos de funciones de activación

La función de activación más común en las redes de neuronas es la función sigmoidal:

$$\varphi(v) = \frac{1}{1 + \exp(-av)}$$

Ecuación 4: Función sigmoidal

En donde a es un parámetro de la función sigmoidal para ajustar la pendiente de la misma, tal y como se puede observar en la siguiente imagen:

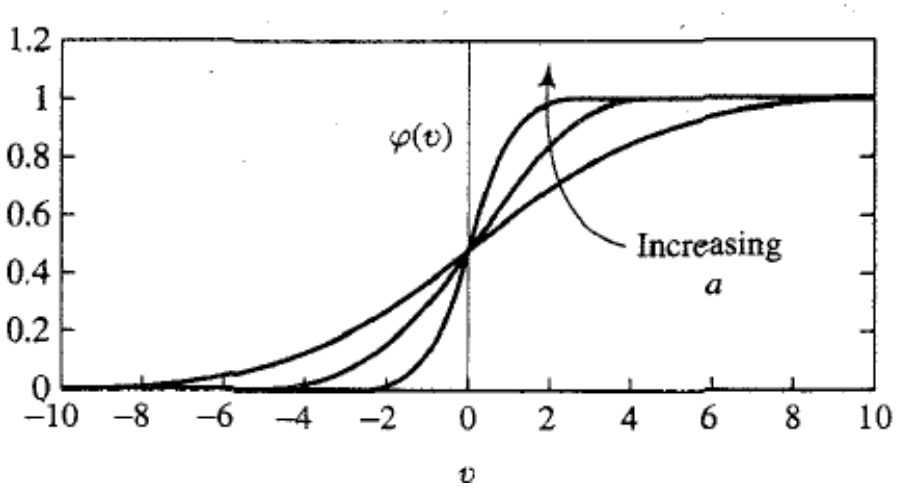


Ilustración 4: Función sigmoidal

Otra funciones típica es la función umbral:

$$\varphi(v) = \begin{cases} 1 \; si \; v \geq 0 \\ 0 \; si \; v < 0 \end{cases}$$

Ecuación 5: Función de umbral

Arquitectura de una red neuronal

Las redes de neuronas se suelen estructurar en forma de capas. Hay tres tipos de capas:

- La capa de entrada: está compuesta por aquellas neuronas que reciben los datos externos y se encargan de transmitir esa información a la capa siguiente.
- Capa oculta de neuronas: recibe la información de la capa de entrada y realiza internamente las operaciones necesarias para modelar el problema a tratar. Al añadir una o más capas a la red neuronal, se consigue que la red realice operaciones más complejas.
- Capa de salida: recibe como valores de entrada las señales de salida de la capa oculta de neuronas. Aquí es donde se produce la respuesta final de la red neuronal, tras haber realizado todo el procesamiento en las anteriores capas.

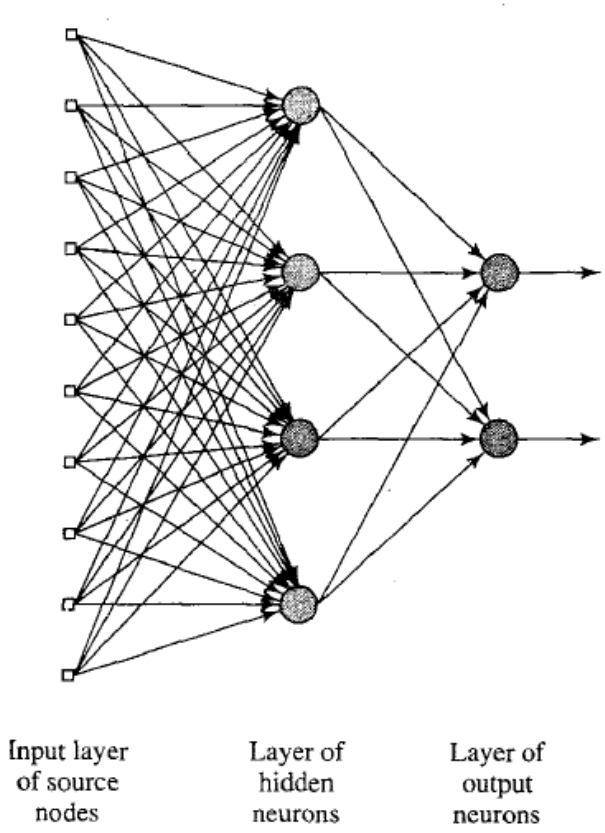


Ilustración 5: Red neuronal multicapa

2.6.2.2 Aprendizaje por backpropagation

Este tipo de aprendizaje, conocida como backpropagation o regla delta generalizada, es el que se aplica a las redes de neuronas multicapa (Ilustración 5). De manera resumida, este proceso consiste en dos fases: una fase hacia delante, en el que la red neuronal recibe los datos de entrada, y va propagando la información capa por capa hasta generar un valor de salida. Este valor de salida se compara con la salida deseada y con ello se calcula el error correspondiente. Tras esta fase, se produce la fase hacia atrás, en donde la información sobre el error obtenido se propaga en sentido contrario, desde la última capa de la red hasta la primera capa. Basándose en el error recibido, se reajustan los pesos de conexión de cada neurona, con el objetivo de obtener una salida más cercana a la deseada, es decir, con el objetivo de que el error disminuya.

A continuación se explica el resumen presentado en (Haykin, 1995) de los pasos que sigue este algoritmo para el entrenamiento de la red neuronal.

Paso 1: Inicialización

Se inicializan los pesos de los enlaces de la red de manera aleatoria.

Paso 2: Presentación de datos de entrenamiento

Se presenta a la red neuronal un conjunto de datos de entrenamiento. Estos datos de entrenamiento consisten en valores de entrada $x_1, x_2, \dots, x_m$. con sus respectivas salidas deseadas $d_1, d_2, \dots, d_m$. Para cada instancia de los datos de entrenamiento, se realizan las operaciones de la fase hacia delante y fase hacia atrás explicados en los puntos 3 y 4 respectivamente.

Paso 3: Computación hacia delante

Siendo cada instancia de datos de entrenamiento un par $(x(n), d(n))$, se computan en cada neurona el valor de salida; estos valores se envían hacia delante a las capas adyacentes, hasta producir una salida final de la red. El valor $v_j^{(l)}(n)$ de una neurona j en la capa l es el valor que se le proporciona a la función de salida:

$$v_j^{(l)}(n) = \sum_{i=0}^{m} w_{ji}^{(l)}(n) y_i^{(l-1)}(n)$$

Ecuación 6: Valor de entrada de la función de salida

en donde $y_i^{(l-1)}(n)$ es el valor de salida de la neurona i en la capa $l-1$ en la iteración n y $w_{ji}^{(l)}(n)$ es el peso de la neurona j en la capa l que recibe la señal de la neurona i en la capa $l-1$. Por lo tanto, el valor de salida de una neurona j en la capa l es:

$$y_i^{(l)} = \varphi(v_j(n))$$

Ecuación 7: Valor de salida de la neurona

Si la neurona j pertenece a la primera capa, entonces:

$$y_i^{(0)} = x_j(n)$$

Ecuación 8: Valor de salida de una neurona en la primera capa

en donde $x_j(n)$ es el elemento número j del vector de entrada $x(n)$.

Si la neurona j pertenece a la capa de salida entonces:

$$y_i^{(L)} = o_j(n)$$

Ecuación 9: Valor de salida de una neurona en la última capa

en donde L indica la profundidad de la red.

Por último, se calcula la señal de error:

$$e_j(n) = d_j(n) - (o_j(n)$$

Ecuación 10: Señal de error

en donde $d_j(n)$ es el elemento número j del vector de salidas deseadas $d(n)$.

Paso 4: Computación hacia atrás

Se calculan los llamados gradientes δ de la red, que se definen de la siguiente manera:

$$\delta_j^{(l)}(n) = \begin{cases} e_j^{(L)}(n)\varphi_j'\left(v_j^{(L)}(n)\right) & para\ la\ neurona\ j\ en\ la\ capa\ de\ salida\ L \\ \varphi_j'\left(v_j^{(l)}(n)\right)\sum_k \delta_k^{(l+1)}(n)w_{kj}^{(l+1)}(n) & para\ la\ neurona\ j\ en\ la\ capa\ oculta\ l \end{cases}$$

Ecuación 11: Gradientes de la red

En donde φ_j' indica la derivada con respecto al argumento. Se ajustan los pesos de los enlaces de la capa l de la siguiente manera:

$$w_{ji}^{(l)}(n+1) = w_{ji}^{(l)}(n) + \alpha\left[w_{ji}^{(l)}(n-1)\right] + \rho\delta_j^{(l)}(n)y_i^{(l-1)}(n)$$

Ecuación 12: Ajuste de pesos

en donde ρ es un ratio de aprendizaje y α una constante.

Paso 5: Iteración

Se repiten los puntos 3 y 4 en los demás ejemplos del conjunto de entrenamiento hasta que el error global obtenido resulta aceptablemente pequeño:

$$E = \frac{1}{2M}\sum_{j=1}^{M} e_j^2(n)$$

Ecuación 13: Error global

en donde $e_j(n)$ indica el error debido a la diferencia de la salida de la red $y_j(n)$ y la salida deseada $d_j(n)$ en j y M indica el número total de instancias de entrenamiento.

2.6.2.3 Deep Learning

En este subapartado se resumirá brevemente la historia de las redes neuronales y cómo éstos se han extendido a lo que se conoce en la actualidad como el aprendizaje profundo o el Deep Learning.

En sus inicios, las redes neuronales no tuvieron mucho éxito en el campo de las Ciencias de la Computación. Posiblemente el primer ejemplo de red neuronal es el Perceptron simple, desarrollado por Rosenblatt (1957). Éste no tenía la capacidad de aproximar muchas

funciones no lineales, como por ejemplo, la función XOR. Para solucionar este problema, surgió la idea de utilizar varias capas en la red neuronal, consiguiendo así aumentar la gama de funciones no lineales que se podían aproximar. Pero este modelo requería de una gran potencia computacional y una enorme cantidad de datos para funcionar óptimamente, por lo que no han tenido éxito hasta esta última década, en la que se ha producido un aumento masivo de datos disponibles, y se ha mejorado la potencia de los ordenadores.

En estos últimos años se ha vuelto muy popular utilizar redes de neuronas con múltiples capas. Gracias a ello, se han conseguido resultados excelentes en áreas como el reconocimiento de voz, reconocimiento de imágenes o reconocimiento del lenguaje (Le, 2015a). El éxito de las redes neuronales con varias capas ha dado como origen el término "Deep Learning". Éste término se utiliza actualmente para referirnos a redes neuronales que utilizan más de dos capas.

La principal razón por la que este tipo de redes han tenido un gran impacto, es que éstas tienen la capacidad de aproximar muchos tipos de funciones no lineales gracias a la manipulación de una gran cantidad de parámetros. Con ello consiguen aproximar problemas muy complejos de manera muy precisa (Le, 2015a). La otra razón de su éxito es su flexibilidad: pueden cambiar fácilmente su estructura con el objetivo de adaptarse a problemas o dominios específicos (Le, 2015b). En el siguiente apartado se explicará de manera breve esta cualidad de las redes de neuronas.

2.6.2.4 Flexibilidad de las redes neuronales

Las redes con arquitecturas profundas destacan por su flexibilidad en comparación con las herramientas tradicionales utilizadas en el aprendizaje automático. Es posible modificar su estructura dependiendo del problema que se quiera resolver. A continuación se describirá las modificaciones más comunes siguiendo a Le (2015b):

- Uso de Autoencoders para el aprendizaje no supervisado y la compresión de datos.
- Uso de redes de neuronas convolucionales para conseguir la invarianza traslacional.
- Predicción de variables que dependen del tiempo mediante redes de neuronas recurrentes.

Autoencoders

Los Autoencoders son un tipo de redes de neuronas que se basan en el aprendizaje no supervisado. Son redes cuya estructura contiene una capa de entrada, una capa de salida, y una o más capas ocultas que conectan las dos anteriores, pero con la particularidad de que el número de neuronas en la capa de entrada y la de salida coinciden. Su objetivo es la de reconstruir la información de entrada X, a diferencia del objetivo tradicional de predecir el valor de Y dado valores de entrada X. Este tipo de redes se utilizan normalmente para problemas como la compresión de datos o la visualización.

Por otra parte, también sirven para pre-entrenar otras redes de neuronas. Tal y como se ha explicado en el aprendizaje backpropagation del apartado 2.6.2.1, normalmente los pesos de los enlaces entre neuronas se inicializan aleatoriamente. Esto no es muy eficiente en el caso de que las redes neuronales tengan arquitecturas profundas, debido a que puede alargar y dificultar la fase de aprendizaje. Por ello, ha surgido la idea de utilizar los autoencoders, con el objetivo de que obtenga previamente información sobre los datos y conseguir así pre-entrenar las redes neuronales inicializando los pesos de una manera más lógica (Bengio, Lamblin, Popovici y Larochelle, 2007).

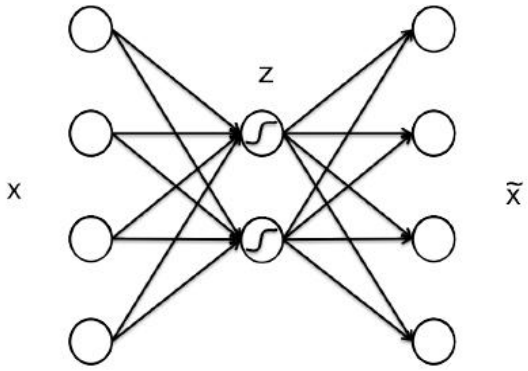


Ilustración 6: Autoencoder

Redes de neuronas convolucionales

Las redes de neuronas convolucionales son los que más éxito han tenido entre los algoritmos de Deep Learning, sobretodo en áreas como el reconocimiento de imágenes (Ciresan et al,.2011).

En las redes neuronales con estructuras clásicas cada neurona está conectada a todas sus neuronas de entrada:

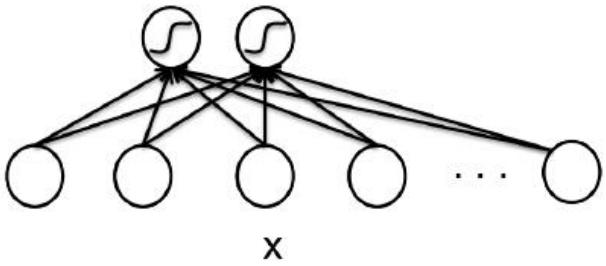


Ilustración 7: Neuronas conectadas con todas las neuronas de entrada

Esta estructura no es muy eficiente cuando el número de entradas es excesivamente grande, debido a que cada neurona tiene demasiadas conexiones. Por ejemplo, si se intenta tratar una imágen de 100x100 píxeles, cada neurona tendrá que lidiar 10.000 parámetros. Para tratar estos problemas de manera más eficiente, se puede que cada

neurona esté conectada solamente con un número fijo de entradas, tal y como se muestra en la siguiente imagen:

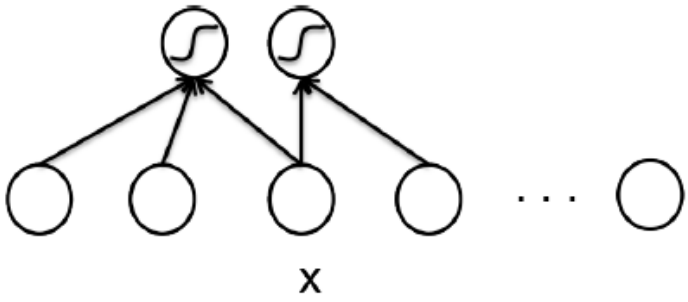


Ilustración 8: Red neuronal convolucional

Este tipo de redes son mucho más eficientes que las redes con arquitecturas tradicionales, y a través de diferentes restricciones que se le apliquen, se pueden obtener propiedades muy interesantes tales como la invarianza traslacional (Le, 2015b).

Redes de neuronas recurrentes

Se trata de un tipo de red neuronal en la que las conexiones entre distintas unidades forman un ciclo. Esta arquitectura permite a la red tener la capacidad de modelar comportamientos dinámicos en la que hay que tener en cuenta la variable del tiempo, por ejemplo, en la predicción del valor de las acciones en una compañía. Este tipo de redes han tenido también bastante éxito en áreas como el reconocimiento de letras o voz (Le, 2015b).

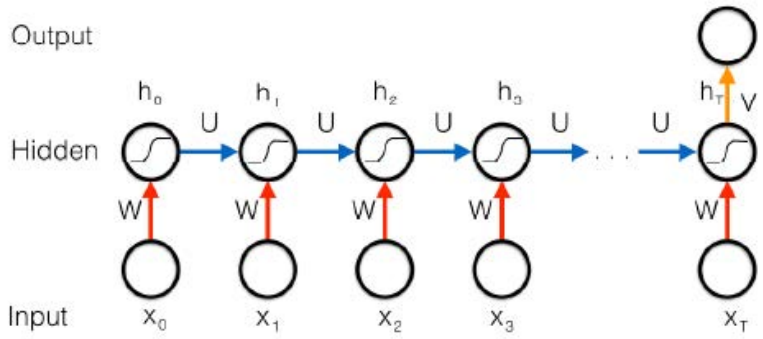


Ilustración 9: Red neuronal recurrente

2.7 Modelo lineal

En este apartado se explicará el modelo lineal siguiendo a Rencher y Schaalje (2007).

Con el modelo lineal se intenta modelar la relación una variable respuesta y, con una o más variables predictoras X. Por ejemplo, en nuestro caso, el coste de un siniestro depende varias variables, tales como la sustitución de sus piezas, horas de mano de obra, etc.

Un modelo lineal que relaciona la variable respuesta y con varios predictores tiene la siguiente forma:

$$\mu = E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

Ecuación 14: Modelo de regresión lineal

Los parámetros $\beta_0, \beta_1, \ldots, \beta_k$ se llaman coeficientes de regresión, e indican de qué manera influye la variable x en la variable respuesta; ε se refiere a una variación aleatoria de y que no está explicada por los regresores x. Esta variación aleatoria surge debido a otras causas que afectan a y además de X, pero que son desconocidas y no se han observado.

Las regresiones lineales tienen varias aplicaciones prácticas (Goldberger, 1991):

1. Predicción del valor de una variable dada una nueva instancia.
2. Descripción y explicación de los datos: se usan los modelos estimados con el objetivo de resumir o describir los datos observados.
3. Estimación de parámetros: sus valores pueden tener implicaciones teóricas para un modelo.
4. Selección de variables: sirve para determinar la importancia de cada predictor para modelar la variación en y. Los predictores que provocan un cambio considerable en y se mantienen, mientras que aquellos predictores que no influyen significativamente se eliminan. Se mide con ello la fuerza de la relación entre y e x.

Una de las condiciones de este modelo es que debe ser lineal en β, pero no es necesario que sea lineal en x. Un ejemplo de un modelo que es lineal en β pero no en x puede ser el siguiente:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2^2 + \beta_3 x_1 x_2 + \varepsilon$$

Ecuación 15: Linealidad en los parámetros

Para estimar β, se usa una muestra de n observaciones de pares (x, y), siendo el modelo en la observación i:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i, \quad i = 1,2, \ldots, n$$

Ecuación 16: Observación i del modelo lineal

Se requiere una serie de supuestos para la correcta formación del modelo:

1. $E(\varepsilon_i) = 0 \; para \; i = 1,2, \ldots, n$
   o equivalentemente, $E(y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$
2. $var(\varepsilon_i) = \sigma^2 \; para \; i = 1,2, \ldots, n$
   o equivalentemente, $var(y_i) = \sigma^2$
3. $cov(\varepsilon_i, \varepsilon_j) = 0 \; para \; todo \; i \neq j$, o equivalentemente, $cov(y_i, y_j) = 0$
4. No existe multicolinealidad exacta, es decir, que no existe ninguna combinación lineal exacta entre las variables explicativas. En caso de existir, el sistema tendría infinitas soluciones. Un ejemplo de combinación exacta de variables sería: $X_1 = 2X_2$

La primera suposición manifiesta que el modelo es correcto, es decir, que todos los x relevantes se han incluido y que el modelo es lineal. La linealidad en los parámetros implica que un cambio unitario en X tiene el mismo efecto sobre Y con independencia del valor inicial de X. La segunda suposición dice que la varianza de y es constante y por ello no depende de x (homocedasticidad condicional). La tercera suposición declara que cada instancia las diferentes y no están correlacionadas unas con otras. Cuando se cumplen estas condiciones, los βs estimados son los de menor varianza entre los estimadores lineales e insesgados (Goldberger, 1991). Si no se cumple alguna de las condiciones, los estimadores son pobres.

2.7.1. Estimación por mínimos cuadrados

Para los parámetros $\beta_0, \beta_1, \ldots, \beta_k$, se buscan estimadores que minimizan la suma de los cuadrados de las desviaciones entre los y observados y los $\hat{y}$ predichos. Se busca entonces $\hat{\beta}_0, \hat{\beta}_1, .., \hat{\beta}_k$ que minimice:

$$\sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_k x_{ik})^2$$

Ecuación 17: Ajuste por mínimos cuadrados

Cabe destacar que con el valor predicho $\hat{y}_i = \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_k x_{ik}$, nos estamos refiriendo a la estimación de $E(y_i)$, no y_i.

Para obtener los estimadores por mínimos cuadrados, que minimizan la anterior ecuación, basta con derivar $\sum_i \hat{\varepsilon}_i^2$ con respecto a $\hat{\beta}_j$ para obtener $k+1$ ecuaciones, resolviéndolos simultáneamente y obteniendo los $\hat{\beta}_j$. Este proceso se resume de forma más compacta en la siguiente ecuación:

Teorema 2.8.1.a. Si $y = X\beta + \varepsilon$, en donde X tiene dimensiones $n\,x\,(k+1)$ y rango $k+1 < n$, entonces el valor de $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, .., \hat{\beta}_k)$ que minimiza la anterior ecuación es:

$$\hat{\beta} = (X'X)^{-1}X'y$$

Ecuación 18: Estimación de los parámetros

Por otra parte, la varianza de la regresión se obtiene de la siguiente manera:

$$\sigma^2 = \frac{1}{n-k-1} \sum_{i=1}^{n} (y_i - x_i' \hat{\beta})^2$$

Ecuación 19: Estimación de la varianza de la regresión

En donde n es el tamaño de la muestra y k es el número de regresores x.

Para medir la bondad de ajuste de una regresión, se utiliza el R-cuadrado:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Ecuación 20: R-cuadrado de una regresión

El R-cuadrado se define como la proporción de la variación de la variable dependiente que es explicada por los regresores.

Y por último, el cálculo de la desviación típica en el estimador $\hat{\beta}_j$ se realiza de la siguiente manera:

$$sd(\hat{\beta}_j) = \frac{\sigma}{\sqrt{nS_j^2(1 - R_j^2)}}$$

Ecuación 21: Estimación del error estándar de los regresores

En donde R_j^2 es el R-cuadrado obtenido al realizar una regresión de x_j sobre las demás variables x, S_j^2 es la varianza muestral del regresor X, y n es el tamaño de la muestra.
Los estimadores de mínimos cuadrados verifican las siguientes propiedades:

- Linealidad en las observaciones de Y.
- Insesgadez en los estimadores.
- Teorema de Gauss-Markov: los parámetros estimados $\hat{\beta}_0, \hat{\beta}_1, .., \hat{\beta}_k$ son los de menor varianza entre los estimadores lineales e insesgados.
- Consistencia.

2.7.2 Evaluación de supuestos

Al obtener una regresión con el modelo lineal, para realizar un análisis correcto de los resultados obtenidos es necesario comprobar que se cumplen las condiciones de multicolinealidad, homocedasticidad e independencia. Para comprobar dichos supuestos, se usan los test del factor de inflación de la varianza, el test de Breusch Pagan y el test de Dubin Watson respectivamente. Por otra parte, también hay que comprobar que la variable dependiente sigue una distribución normal. El estudio de distribuciones se realizará mediante el test de Kolmogorov-Smirnov.

2.7.2.1 Test del factor de inflación de la varianza

El test del factor de inflación de la varianza (VIF) se utiliza para medir la multicolinealidad entre las variables predictoras (Belsley, 1991). Se describe matemáticamente de la siguiente manera:

$$VIF_j = \frac{1}{1 - R_j^2}$$

Ecuación 22: Factor de inflación de la varianza

en donde R_j^2 es el estadístico R^2 de una regresión en X_j sobre los demás regresores, sin utilizar la variable Y. Se puede observar en esta ecuación que cuanto mayor sea R_j^2, mayor será el VIF de ese regresor. Dicho en otras palabras, cuanto mejor se explique el regresor X_j a través de los demás regresores, mayor será su VIF. No existe un valor crítico bien definido que indique que el VIF es alto, pero algunos autores indican que un factor de inflación de varianza mayor que 10 es lo suficientemente grande para indicar que hay un problema de multicolinealidad (Chatterjee y Price, 1991).

Si una multicolinealidad alta, éste incrementa la varianza de los parámetros estimados y aumenta su sensibilidad a pequeños cambiso en el modelo. En otras palabras, se obtienen resultados inestables y difíciles de interpretar (Chatterjee y Price, 1991).

2.7.2.2 Test de Breusch Pagan

Este test se usa para detectar la heterocedasticidad en un modelo de regresión lineal (Breusch y Pagan, 1979). Para este test se utiliza el siguiente contraste de hipótesis:

$$H_0 = el\ modelo\ es\ homocedástico$$

$$H_1 = el\ modelo\ es\ heterocedástico$$

Para calcular el estadístico de Breusch Pagan, se utiliza una regresión de los residuos obtenidos y las variables independientes:

$$e_i^2 = \tau_1 + \tau_2 + \cdots + \tau_k$$

Ecuación 23: Regresión en los errores

Se obtiene de esta regresión el estadístico R^2 y con ello se calcula el estadístico del test de Breusch Pagan:

$$BP = nR^2$$

Ecuación 24: Estadístico de Breusch Pagan

en donde n es el número de observaciones. Este estadístico sigue una distribución de chi-cuadrado con $k - 1$ grados de libertad.

Si no se cumple este supuesto, los estimadores por mínimos cuadrados siguen siendo insesgados y consistentes, pero los errores estándar de las estimaciones son sesgados e inconsistentes, por lo que no se pueden utilizar para hacer inferencia (Goldberger, 1991) En resumen, los estadísticos obtenidos por mínimos cuadrados ya no son el mejor estimado lineal e insesgado.

Si el modelo falla solamente en la heterocedasticidad, es posible arreglar los errores estándar de las estimaciones utilizando los errores robustos a la heterocedasticidad, también conocidos por errores estándar Huber-White (Goldberger, 1991):

$$Var(\beta) = (X^{'}X)^{-1}X^{'}\varphi(X^{'}X)^{-1}$$

Ecuación 25: Error robusto a la heterocedasticidad

En donde $\varphi = \begin{bmatrix} \hat{e}_1^2 & 0 & 0 \\ 0 & \hat{e}_2^2 & 0 \\ 0 & 0 & \hat{e}_n^2 \end{bmatrix}$, es decir, una matriz con los errores obtenidos en la diagonal.

2.8.2.3 Test de Dubin Watson

Este test se utiliza para detectar la autocorrelación entre residuos en un modelo de regresión lineal. (Durbin y Watson, 1950). Para este test se considera el siguiente contraste de hipótesis:

$$H_0 = no\ existe\ autocorrelación$$

$$H_1 = existe\ autocorrelación$$

Para ello se utiliza el siguiente estadístico:

$$d = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

Ecuación 26: Estadístico de Dubin Watson

que se contrastará frente a valores críticos del 1-5%. La implementación en el lenguaje R muestra directamente el p-valor. Sin embargo, si se quiere buscar el valor crítico de estos estadísticos, éstos se pueden encontrar en (Savin y White, 1977).

Si no se cumple este supuesto, los estimadores de mínimos cuadrados no son el mejor método de estimación, se subestimará la varianza calculada y los estadísticos calculados serán demasiado optimistas. En resumen, los parámetros estimados serán ineficientes (Goldberger, 1991).

2.7.3 Modelo Lineal Generalizado

A continuación, se explicarán los modelos lineales generalizados siguiendo a (McCullagh y Nelder, 1983).

Durante las últimas décadas, los LM se han extendido a los modelos lineales generalizados (GLM) debido a dos razones: primero, no se puede asumir que los datos tengan una distribución Normal, y segundo, la media no tiene por qué ser una combinación lineal de las variables explicativas. Los GLM engloban las clásicas regresiones lineales y los modelos ANOVA, regresiones logísticas y modelos probit para variables discretas, y una amplia

familia de modelos para datos categóricos y modelos de Poisson. Todos estos modelos comparten características comunes, como la linealidad en sus parámetros, lo que permite estudiar los modelos lineales generalizados en una sola clase, en vez de tener una serie de modelos no relacionados entre sí. Esta expansión de los LM tradicionales fue introducida por primera vez en (Nelder y Wedderburn, 1972).

Supuestos del GLM

Para una correcta especificación de los modelos lineales generalizados, éstos deben cumplir una serie de supuestos, aunque son menos restrictivos que los modelos lineales.

- Independencia en las observaciones. Se excluyen por ello los datos que muestran autocorrelaciones, por ejemplo, las series temporales.
- Solamente hay un término de error en el modelo. Se excluyen todos los modelos que tengan más de un término de error.
- Ya no se requiere la normalidad y una variancia constante, tal y como sucede con las regresiones lineales.

Generalización del modelo lineal

Para simplificar la transición del LM al GLM, reintroduciremos el modelo lineal (1) en tres partes:

1. El componente aleatorio: los componentes de Y tienen distribuciones normales con $E(Y) = \mu$ y varianza constante σ^2;
2. Componente sistemática: las variables explicativas $x_1, x_2, \dots, x_n$ producen un predictor lineal τ

$$\tau = \sum_{i=1}^{k} x_i \beta_i$$

Ecuación 27: Componente sistemática

3. Un enlace entre el componente sistemático y el componente aleatorio:

$$\mu = \tau$$

Ecuación 28: Función de enlace en el modelo lineal

Esta generalización introduce un nuevo símbolo τ para el predictor lineal. El tercer componente especifica que en el caso del modelo lineal, μ y τ son idénticos. Si escribimos:

$$\tau_i = g(\mu_i)$$

Ecuación 29: Función de enlace

entonces $g(.)$ lo llamaremos la función de enlace. Los modelos lineales generalizados permiten dos extensiones: permite primero que la distribución del primer componente provenga de la familia de distribuciones exponenciales, en vez de poder seguir solamente la distribución Normal; y segundo, la función de enlace del tercer componente se puede transformar en cualquier función monótona y diferenciable.

La familia exponencial de distribuciones tiene la siguiente forma

$$f_Y(y;\theta,\varphi) = exp\left\{\frac{y\theta - b(\theta)}{a(\varphi)} + c(y,\varphi)\right\}$$

Ecuación 30: Familia exponencial de distribuciones

para las funciones específicas a(.), b(.) y c(.).

Algunos ejemplos de distribuciones que pertenecen a la familia exponencial son la Normal, Poisson, Binomial, Gamma, Gaussiana Inversa, etc.

Función de enlace

Como se ha dicho anteriormente, la función de enlace relaciona el predictor lineal τ al valor esperado $\mu = E(Y)$. En el modelo lineal, la media y el predictor lineal son idénticos. Sin embargo, cuando se usan otro tipo de distribuciones como por ejemplo la binomial, se tiene que cumplir la condición $0 < \mu < 1$, y la función de enlace se debe encargar de satisfacer la condición de que los valores estén comprendidos en el intervalo (0,1).

Existen las funciones de enlace canónicas, que se tratan de funciones de enlace especiales para los que existe un estadístico suficiente con una dimensión similar a β en el predictor lineal $\tau = \sum x_i\beta_i$; esto ocurre cuando $\theta = \tau$, en donde θ es el parámetro canónico.

A continuación se muestra las funciones de enlace canónicas pertenecientes a las distribuciones más típicas:

- Normal: $\tau = \mu$
- Gamma: $\tau = \mu^{-1}$

- Poisson: $\tau = \log(\mu)$
- Binomial: $\tau = \log\{\frac{\mu}{1-\mu}\}$
- Gaussiana inversa: $\tau = \mu^{-2}$

2.7.4 Test de Kolmogorov-Smirnov

Es de gran importancia saber qué distribución siguen los datos para saber qué modelo de regresión utilizar, por lo que previamente es necesario realizar un contraste de bondad de ajuste para ver cuál es la distribución estadística que mejor se ajusta a los datos. Con normalidad se puede utilizar un modelo lineal de manera óptima, sin ella, habrá que optar por un modelo lineal generalizado. Uno de los tests más usados es el test de Kolmogorov-Smirnov, en el que se compara la distancia vertical entre la distribución de probabilidad de una distribución conocida, y una distribución empírica (Wilcox, 2005):

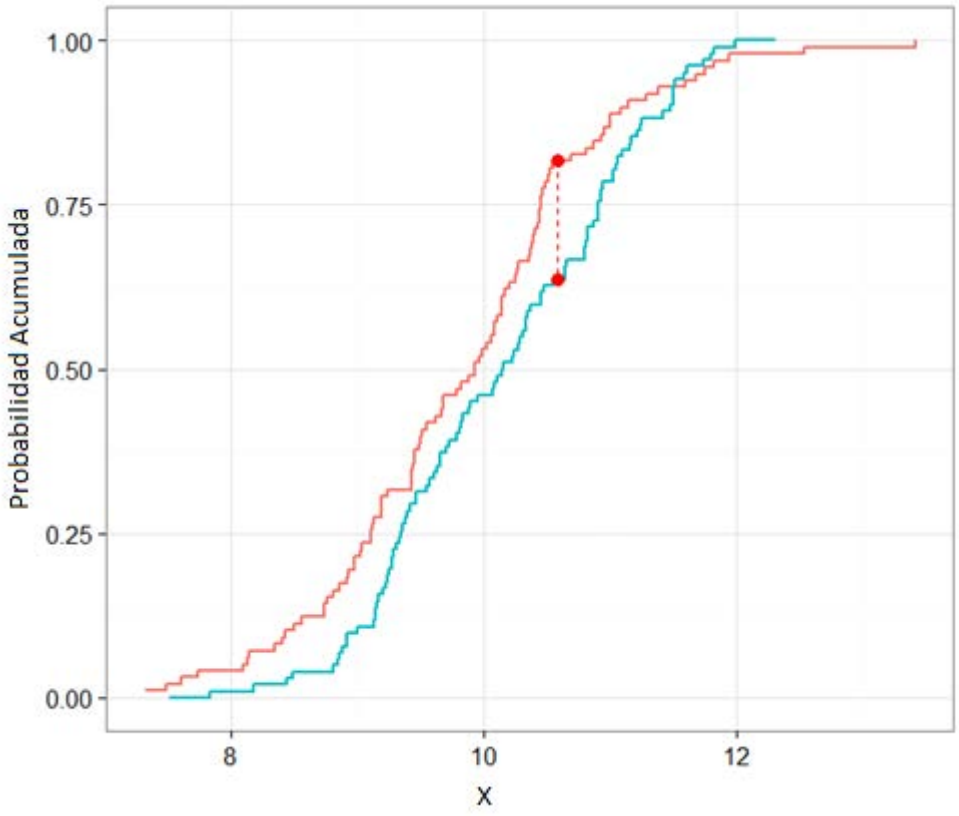


Ilustración 11: Test de Kolmogorov-Smirnov

2.8 Estudios previos y proyectos similares

Se ha realizado una búsqueda en el repositorio de documentos de la Universidad Carlos III de Madrid con el objetivo de encontrar estudios relacionados con el tema que se va a tratar en este proyecto.

Se han encontrado dos trabajos fin de grado que trataban sobre el mismo tema: (González, 2015) y (Anca, 2015). En ambos trabajos fin de grado, se realiza la separación de los datos en grupos basados en el coste económico de la reparación del siniestro y en las zonas de impacto

de cada siniestro con el fin de automatizar el proceso de tasación y facilitar la detección del fraude. El objetivo de estos proyectos es bastante similar al de este estudio: la de crear modelos para poder mejorar la tasación del seguro e implementar mecanismos para detectar el fraude en el seguro de automóviles.

En el caso de E. González y G. Anca, trabajan solamente en una pequeña porción de los datos de las que se dispone en este proyecto. Este proyecto ampliará el alcance a doce segmentos de automóviles y por otra parte, se realizará un estudio distinto a los dos anteriores: en vez de separar los siniestros en diferentes grupos, se realizarán modelos predictivos para las diferentes variables de interés.

Este proyecto complementará todo lo que E. González y G. Anca realizan en sus trabajos. De esta manera, las aseguradoras dispondrán de un kit completo de herramientas para mejorar su sistema de tarificación de seguros.

3. Diseño de la solución

En este apartado se explicarán las decisiones que se han tomado para la realización de este proyecto. Se hablará de todo el proceso llevado a cabo, las herramientas utilizadas (software y hardware) y sus alternativas y por último se explicará la implementación de los algoritmos realizada para llevar a cabo este estudio.

3.1 Herramientas utilizadas

3.1.1 H2O

La herramienta más utilizada en este proyecto es el entorno de desarrollo de algoritmos de aprendizaje automático H2O.

H2O es una plataforma de aprendizaje automático y análisis predictivo que destaca tanto por la rapidez y escalabilidad de sus algoritmos, como por ser una librería de código abierto. Permite construir modelos para el análisis dirigido al Big Data y está enfocada a la producción de las mismas en el entorno empresarial. Contiene los algoritmos más usados actualmente en la rama de aprendizaje automático tales como Deep Learning, árboles de decisión o GLRM, lo que hace a H2O una de las plataformas más usadas del momento para el análisis de datos.

Se puede acceder a esta librería mediante diversos lenguajes de programación como R, Python Java o Scala, y ésta puede ser ejecutado eficientemente tanto en tu ordenador personal como en servidores, en los sistemas operativos más usados del momento (Windows, Mac o Linux). Ofrece otras características tales como la rapidez de la recogida de datos distribuidos y de su transformación, y ofrece otros aspectos importantes para el entorno empresarial tales como la seguridad o la autenticación.

El código de esta librería está escrito principalmente en el lenguaje Java, lo que permite un manejo rápido y eficiente de grandes cantidades de datos. Por otra parte, destaca por la lectura de sus datos en paralelo, su distribución en clusters y su almacenamiento de manera comprimida. Todo esto ha permitido que los algoritmos de esta librería se hayan implementado de una manera eficiente, lo que se traduce en la creación de modelos más precisos, rápidos y con mejores predicciones y por la facilidad de creación de las mismas.

La mejor alternativa a H2O es Tensorflow, otra librería de código abierto para el aprendizaje automático usado en Google que se publicó en noviembre del 2015. Se ha elegido la plataforma H2O frente a la de Google debido a que ésta última, ´ha sido publicada recientemente y por ello no ofrece tanta documentación como H2O.

Otras alternativas serían Theano, Caffe o Torch, pero éstas o tienen mayor dificultad de implementación, o sus ejecuciones son menos eficientes, por lo que tampoco han sido consideradas en este proyecto.

3.1.2 RStudio

Se ha elegido esta plataforma de desarrollo para R ya que ofrece un entorno de desarrollo integrado (IDE) que facilita el desarrollo de aplicaciones al programador. Ofrece una plataforma más visual e intuitiva que la consola de R estándar. Tiene herramientas que mejoran la eficiencia en el trabajo, tales como la posibilidad de buscar documentación de ayuda directamente en una consola, en vez de tener que buscarlo en un navegador web; o un sistema de gestión de gráficas, en la que en vez de mostrar ventanas de imágenes solapadas (como aparece en R), te muestra solamente un panel en donde puedes elegir qué gráfica mostrar.

Por otra parte ofrece un espacio de trabajo en donde se pueden ver todas las variables de interés, lo que facilita la tarea del programador y la depuración del código.

3.1.3 Otro software

Referente al software, se han utilizado otras herramientas como Microsoft Excel 2015 y Microsoft Word 2015.

La primera ha servido para crear tablas en la que se realizan las estadísticas y comparaciones entre los modelos desarrollados, a su vez de gráficas para mostrar los resultados visualmente. Por otra parte también se ha utilizado para ver las características de los ficheros CSV proporcionados, en la que se encuentran los datos de los siniestros.

Para la redacción del presente documento, se ha usado el programa Microsoft Word 2015, ya que es el editor de texto más utilizado, tiene mucha documentación y permite la redacción y estructuración del documento de una manera muy intuitiva.

3.1.4 Sistemas Operativos

Todo el proyecto se ha realizado en Windows 10, ya que se las herramientas de programación utilizadas soportaban el uso en este sistema operativo.

3.2 Descripción de los datos

Se dispone de 12 ficheros previamente procesados a partir de una base de datos de siniestros en España de una importante compañía de seguros. Cada fichero se corresponde a un tipo de vehículo según variables clasificadoras como la marca, modelo y tamaño. Estos datos han sido tomados durante un periodo de 5 años y conjuntamente cuentan con un total de 4.147.715 de partes de accidente. A continuación se muestra el número total de siniestros por segmento:

Segmento	Número de siniestros
A	161.698
B	712.552
C	865.814
D	673.479
E	270.407
F	42.058
G	652.025
H	124.165
TA	215.130
TB	172.741
TC	171.513

Tabla 1: Número de siniestros por segmento

Los datos están guardados en ficheros con formato CSV, que son documentos en donde se representan los datos en forma de tabla, las columnas están separadas por comas y representan las variables de un siniestro (zona afectada y operación realizada), cada línea o fila representa un siniestro y la primera fila de todas indica el nombre de las variables del siniestro.

Cada siniestro está representado por 87 atributos. A continuación se describe detalladamente la información que aporta cada atributo:

- Coste total del siniestro (sin IVA).
- Coste total de las nuevas piezas.
- Horas totales de mano de obra en pintura.
- Horas totales de mano de obra en chapa.
- Número de piezas sustituidas en cada zona.
- Número de piezas reparadas en cada zona.
- Número de piezas pintadas en cada zona.
- Número de secuencia (número identificativo del siniestro).
- Modelo del automóvil (a qué segmento pertenece).

El automóvil se ha dividido en 27 zonas:

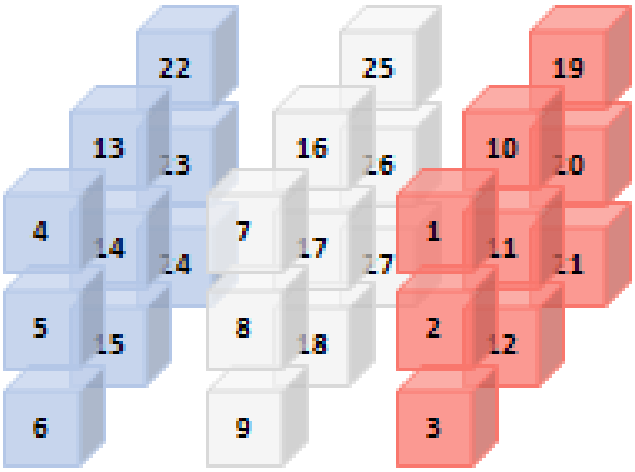


Ilustración 10: Automóvil separado en 27 zonas

En donde los cubos rojos hacen referencia a un lateral izquierdo del automóvil, los cubos azules el lateral derecho, las zonas 1-9 el frente del automóvil, y las zonas 19-27 el trasero.

3.3 Descripción del procesado de datos

En esta sección se explicarán las operaciones y estudios de los datos que se han tenido que realizar para conseguir el funcionamiento más correcto y óptimo de los algoritmos empleados.

3.3.1 Normalización

Debido a la gran cantidad de datos y al gran rango que puede existir entre el valor mínimo y el valor máximo en los atributos, es de conveniencia normalizar los datos.

Normalizar consiste en limitar el rango total de un atributo entre 0 y 1, transformando el valor de cada instancia a su valor proporcional en dicho rango. La fórmula que se usa para normalizar los datos es la siguiente:

$$x_{nuevo} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Ecuación 31: Normalización

Posteriormente, para volver a poner los datos en la escala original, simplemente hay que desnormalizar los datos:

$$x_{original} = (\max(x) - \min(x)) * x + \min(x)$$

Ecuación 32: Desnormalización

La normalización es bastante importante dado que existen datos en diferentes escalas. Por otra parte, permite un aprendizaje más rápido y eficiente en las redes de neuronas, y reduce la posibilidad de que la solución se estanque en un mínimo local.

3.3.2 Aleatorización

Es conveniente presentar los datos de manera aleatoria para los algoritmos de minería de datos, para evitar problemas de sobreajuste de datos y evitar sesgos en el aprendizaje. También, para comprobar la eficiencia de un algoritmo, conviene realizar varias ejecuciones en conjuntos de datos diferentes (aleatorios) para confirmar que el modelo tiene la capacidad de generalizar sus predicciones.

3.3.3 Filtrado de atributos

Dependiendo de la prueba que se está realizando, resulta conveniente ignorar unos atributos y mantener otros. Por otra parte también, dependiendo del segmento hay zonas a las que no se han realizado ningún tipo de reparación (sustitución, reparación, pintura), por lo que se han eliminado del fichero dado que no es conveniente proporcionar información no útil al algoritmo.

3.3.4 División de datos

3.3.4.1 División de los datos para el estudio

Existe un gran rango en los valores a predecir: por ejemplo en el coste, el máximo puede rondar las decenas de miles, mientras que el mínimo puede estar en el orden de las centenas. Estudiando los datos, se ha descubierto que por ejemplo en el coste total, el tercer cuartil suele estar alrededor de los 1.000€, mientras que el máximo puede llegar hasta unos 20.000€. Esta gran diferencia implica tener predicciones más inexactas, ya que los modelos solamente aprenden a predecir los casos más comunes (en este caso para los valores pequeños hasta el 3º cuartil) y no es capaz de aprender los casos "extremos".

Por lo tanto, con el objetivo de tener predicciones más precisas, se han creado primero modelos utilizando todos los datos, y luego otros modelos utilizando solamente el subconjunto entre el mínimo al 3º cuartil. En este proyecto no se ha considerado estudiar el subconjunto 3º-4º cuartil, dado que éstos se tratan de siniestros complicadísimos y únicos,que conllevan costes gigantescos, y suelen ser pocos casos comparado con el primer subconjunto, por lo que no resulta de mucha utilidad para las empresas aseguradoras analizar esta parte de los datos

3.3.4.2 División de datos de entrenamiento y de prueba

Para comprobar la eficiencia del análisis predictivo de un modelo, se suelen dividir los datos disponibles en dos subconjuntos disjuntos: el subconjunto de entrenamiento (training set) y el subconjunto de prueba (test set). El algoritmo usará el subconjunto de entrenamiento para aprender las características de los datos y sacar un modelo capaz de predecir la solución objetiva, y posteriormente se usará el subconjunto de prueba para ver si el modelo tiene la capacidad de generalizar nuevos casos que no se ha visto en los datos de entrenamiento. Una

las proporciones más comunes usadas en el análisis de datos es tener el 70% de datos para el entrenamiento y el 30% para el test, que es la que se ha utilizado en este proyecto.

3.4 Algoritmos de regresión

En esta sección se explicará el proceso que se ha seguido para obtener los modelos de predicción en las variables de interés de los siniestros de automóviles. Como se ha explicado en el apartado 1.2 de este documento, se han usado dos algoritmos distintos provenientes de ramas distintas (inteligencia artificial y estadística) con el objetivo de comparar el rendimiento de ambos: las redes de neuronas y el modelo lineal.

3.4.1 Predicción con redes neuronales

Para la ejecución de este algoritmo, H2O requiere que se especifiquen los siguientes parámetros:

- *Datos de entrenamiento*: el subconjunto de datos usado para construir el modelo. Se obtiene a partir de los datos de los ficheros CSV. Se ha cogido de manera aleatoria un subconjunto del 70% de los datos. Se debe eliminar cualquier atributo que no se vaya a usar.
- *Datos de test*: el subconjunto de datos usado para testear la eficiencia y precisión del modelo. Se trata del otro 30% de los datos de los ficheros CSV.
- *Epochs*: el número de veces que se recorre la base de datos para entrenarse.. Para un correcto aprendizaje, se necesita recorrer los datos varias veces. Cuanto mayor sea la cantidad de datos a analizar, mayor epochs se necesitará. Hay que tener cuidado con poner un número excesivamente grande ya que puede llevar al sobreajuste de datos. Éste se ha ajustado según la cantidad de datos a analizar.
- *Función de activación*: función de activación de las neuronas. Las alternativas que había eran las siguientes funciones:
  - Tanh: $f(\alpha) = \frac{e^{\alpha} - e^{-\alpha}}{e^{\alpha} + e^{-\alpha}}$
  - Rectifier: $f(x) = \max(0, \alpha)$

  La primera permite que la fase de entrenamiento del algoritmo converja más rápido, gracias a la simetría que ofrece alrededor de 0 (Candel, Parmar, Ledell, y Arora, 2016). En la segunda, se suelen obtener empíricamente mejores resultados y tiene mejores rendimientos en determinados problemas, por lo que se ha elegido ésta.
- *Número de capas y neuronas*: aquí se especifican el número de neuronas y de capas ocultas. Por ejemplo, si se pone (100, 100), quiere decir que la red neuronal tiene dos capas ocultas de 100 neuronas cada una y en total tendría: 1 capa de entrada con tantas neuronas como entradas haya, dos capas ocultas de 100 neuronas cada una, y una capa de salida en donde se calcula la solución final. Se ha seguido una estructura de 4 capas ocultas con una arquitectura autoencoder: la primera capa tendrá tantas neuronas como entradas haya, la segunda y tercera capa la mitad de la primera, y la última las mismas neuronas que la primera. Es decir, siendo n el número de entradas, se le especifica al algoritmo: $(n, \frac{n}{2}, \frac{n}{2}, n)$

- *Sparse*: se trata de una variable booleana. Si se activa se "dispersan" lo cual mejora la eficiencia del manejo de datos cuando se tiene un gran número de valores que son 0. Esto interesa activarlo porque en el caso de los siniestros, hay bastantes zonas que no han sido afectadas, dando lugar a muchos atributos con valores nulos.

3.4.2 Predicción mediante el modelo lineal generalizado

Para la ejecución de este algoritmo en R, se tienen que especificar los siguientes parámetros:

- *Fórmula a usar*: se refiere una descripción del modelo a ajustar. Generalmente se especifica de la siguiente manera: Coste ~ Sustitución + Pintura + Reparación. Esto indica en realidad que Coste es la variable a predecir (Y), Sustitución, Pintura y Reparación son las variables predictoras (X) y se está ajustando el modelo: $Y = \beta_1 Sustitución + \beta_2 Pintura + \beta_3 Reparación + \beta_0 + \varepsilon$
- *Familia del modelo:* se refiere a la distribución que siguen los datos del modelo, por ejemplo distribución Gamma, Normal, Poisson, etc. Esto determinará la función de enlace a usar en el modelo glm. Cabe destacar que si se elige una distribución normal, el modelo utilizado es un modelo lineal.
- *Datos:* el equivalente a los datos de entrenamiento de las redes de neuronas. Provienen del archivo CSV que contiene los datos de los siniestros. Al igual que antes se ha seguido una proporción estándar de usar el 70% de datos para entrenamiento, y el 30% para el test de su eficiencia.
- *Valor inicial de los parámetros:* por ejemplo, en la distribución Gamma, los parámetros tienen que ser igual o mayores que 0.

4. Evaluación de los resultados

En este apartado se mostrarán todos los resultados obtenidos con las redes de neuronas y el LM, y con ello se realizará un estudio sobre las características esenciales de cada segmento de automóviles y a su vez se hará una comparación de los modelos utilizados.

Cabe mencionar que para determinar la tasa de acierto de los modelos y compararlos entre ellos, se ha utilizado el error obtenido en cada predicción del conjunto de datos de test, mediante el error absoluto:

$$Error(x, y) = |x - y|$$

Ecuación 33: Error absoluto

Con ello se ha calculado el error medio por predicción:

$$Error\ medio = E(Error\ absoluto) = \frac{1}{N}\sum_{i=1}^{N} error_i$$

Ecuación 34: Error medio

Y para determinar la precisión del modelo, se comparará el error medio por predicción con el valor medio de la variable dependiente. Con ello se sabrá cuánto difiere de media los valores predichos y los valores reales.

$$Precisión = 1 - \frac{Error\ medio}{E(Y)}$$

Ecuación 35: Precisión

en donde $E(Y)$ es la media del coste total, coste de piezas, horas de pintura o horas de chapa en el segmento.

También es importante considerar la desviación típica del error. Interesa que la desviación sea lo más pequeño posible, ya que valores grandes de desviación disminuye la precisión de la información predicha. Por ejemplo, si tienes una desviación típica muy baja con un error medio de 100€, sabes que el valor predicho y_i siempre estará 100€ por encima o por debajo, pero en el caso tener una desviación típica alta, no sabrás si la predicción es muy buena, o si te has equivocado en un valor mucho mayor que 100€. Dicho en otras palabras, el error en la predicción será mucho más volátil y por ello más difícil de interpretar.

$$Desviación\ típica = \frac{\sum_{i=1}^{n}(x_i - E(X))^2}{n-1}$$

Ecuación 36: Desviación típica

4.1 Análisis descriptivo de los datos

Antes de mostrar los resultados obtenidos, es imprescindible analizar con qué datos nos estamos enfrentando. A continuación se en diferentes tablas estadísticos descriptivos de las variables a predecir.

Coste total

Coste Total	A	B	C	D	E	F
Mínimo	0,01 €	0,01 €	0,01 €	0,01 €	0,01 €	0,01 €
1º cuartil	249,00 €	274,40 €	298,60 €	324,40 €	356,40 €	425,00 €
Mediana	437,00 €	484,90 €	515,30 €	550,00 €	616,50 €	783,90 €
Media	716,00 €	819,00 €	902,50 €	952,80 €	1.144,00 €	1.421,00 €
3º cuartil	851,20 €	998,00 €	1.159,00 €	1.159,00 €	1.304,00 €	1.584,00 €
Máximo	10.330,00 €	12.020,00 €	15.290,00 €	15.290,00 €	20.140,00 €	21.550,00 €
Desviación típica	844,78 €	991,23 €	1.145,21 €	1.174,87 €	1.592,59 €	1.972,35 €

Coste Total	G	H	S	TA	TB	TC
Mínimo	0,01 €	0,01 €	0,01 €	0,01 €	0,01 €	0,01 €
1º cuartil	332,10 €	345,50 €	340,00 €	331,00 €	342,10 €	411,00 €
Mediana	581,80 €	614,00 €	617,10 €	575,70 €	601,70 €	726,30 €
Media	938,80 €	981,80 €	1.175,00 €	9.170,10 €	1.031,00 €	1.276,00 €
3º cuartil	1.181,00 €	1.200,00 €	1.316,00 €	1.147,00 €	1.205,00 €	1.473,00 €
Máximo	13.940,00 €	15.630,00 €	21.120,00 €	17.420,00 €	19.410,00 €	22.040,00 €
Desviación típica	1.074,49 €	1.162,65 €	1.715,62 €	1.262,07 €	1.397,00 €	1.740,10 €

Tabla 2: Estadísticos descriptivos del coste total de los siniestros

En la tabla 2 se muestran diferentes estadísticos para describir el coste total. Se puede observar que los datos son muy heterogéneos. El coste medio de reparación de un automóvil oscila desde los 580€ hasta los 1420€ dependiendo del segmento. Los costes totales más altos se encuentran en los segmentos F, TC y S, mientras que los más bajos se encuentran en los segmentos A, B y C.

Los valores oscilan entre los 0,01€ hasta los 10.000-20.000€. . Se puede observar que la mediana tiene valores entre los 400-800€, lo que indica que la mayoría de los costes están alrededor de esos valores. Esto se corrobora con el valor del tercer cuartil, que oscila entre los 800 y 1500€. Lo que más destaca es la diferencia que hay entre el 3º cuartil y el valor máximo. Por ejemplo, en el segmento A, el tercer cuartil es 851,20€ y el máximo es de 10.330€, es decir, más de 10 veces que el primer valor. Lo mismo sucede en los demás segmentos, en donde la diferencia es más abismal todavía. Esto indica que hay un grupo reducido de siniestros muy complicados que han supuesto unos costes de reparación enormes.

Por otra parte, la desviación típica es bastante alta, oscilando entre valores de 850-1700€, y llegando a ser incluso mayor que la media. Esto es un indicador de la gran dispersión que hay en los datos, originada sobre todo por el conjunto de datos superiores al 3º cuartil.

Coste de piezas

Coste Piezas	A	B	C	D	E	F
Mínimo	0,00 €	0,00 €	0,00 €	0,00 €	0,00 €	0,00 €
1º cuartil	75,95 €	59,14 €	65,15 €	75,66 €	89,49 €	120,10 €
Mediana	186,80 €	202,60 €	246,00 €	268,80 €	336,40 €	450,00 €
Media	378,50 €	438,40 €	516,40 €	534,30 €	728,20 €	961,40 €
3º cuartil	389,10 €	445,60 €	510,50 €	565,20 €	726,30 €	980,70 €
Máximo	7.300,00 €	8.907,00 €	10.890,00 €	11.300,00 €	15.930,00 €	18.150,00 €
Desviación típica	614,45 €	765,91 €	937,46 €	931,31 €	1.352,62 €	1.732,72 €

Coste Piezas	G	H	S	TA	TB	TC
Mínimo	0,00 €	0,00 €	0,00 €	0,00 €	0,00 €	0,00 €
1º cuartil	75,06 €	84,74 €	73,88 €	100,40 €	97,81 €	123,30 €
Mediana	236,90 €	281,10 €	311,60 €	308,20 €	335,20 €	409,20 €
Media	478,70 €	541,40 €	764,80 €	603,20 €	639,40 €	820,70 €
3º cuartil	521,00 €	599,70 €	746,30 €	615,00 €	670,20 €	847,70 €
Máximo	10.490,00 €	12.540,00 €	16.650,00 €	13.810,00 €	15.360,00 €	17.480,00 €
Desviación típica	819,07 €	926,84 €	1.469,92 €	1.060,55 €	1.160,08 €	1.487,14 €

Tabla 3: Estadísticos descriptivos del coste de piezas de los siniestros

En la tabla 3 se muestran los estadísticos para describir el coste de piezas de los siniestros. Al igual que sucede en el coste total, los datos son muy diferentes entre los segmentos. El coste medio de piezas requeridas oscila desde los 186€ hasta los 961€ dependiendo del segmento. Los costes de piezas más altos se encuentran en los segmentos F y TC, mientras que los más bajos se encuentran en los segmentos A y B.

Los valores oscilan entre los 0,00€ hasta los 7300-18150€. La mediana tiene valores entre los 186,8-450€, lo que indica que la mayoría de los costes están alrededor de esos valores. Aquí también destaca la diferencia que hay entre el 3º cuartil y el valor máximo. Por ejemplo, en el segmento A, el tercer cuartil es 389,10€ y el máximo es de 7.300€, el cual es casi 20 veces el primer valor.

Se puede observar que aquí también hay desviaciones típicas muy altas, con valores entre los 600-1700€, fruto del gran rango de valores en el coste de piezas.

Horas de pintura

Horas Pintura	A	B	C	D	E	F	G	H	S	TA	TB	TC
Mínimo	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
1º cuartil	2,50	2,72	2,65	2,70	2,81	2,85	2,96	3,06	2,67	2,70	2,59	2,66
Mediana	3,82	4,20	4,25	4,42	4,60	4,72	4,60	4,83	4,40	4,30	4,26	4,50
Media	4,58	5,30	5,63	6,07	6,27	6,64	6,09	6,50	5,87	5,71	5,99	6,30
3º cuartil	5,45	6,18	6,55	7,04	7,38	7,74	7,00	7,40	7,00	6,55	6,95	7,46
Máximo	30,50	39,65	44,00	47,00	46,55	53,49	40,00	41,11	44,23	36,84	48,10	45,90
Desviación típica	3,28	3,99	4,54	5,07	5,23	5,80	4,85	5,38	4,73	4,65	5,17	5,41

Tabla 4: Estadísticos descriptivos de las horas de pintura de los siniestros

En la tabla 4 se muestran los estadísticos para describir las horas de mano de obra en pintura de los siniestros. Se puede observar que las horas medias de pintura requeridas oscilan entre las 4,6-6,65 horas, dependiendo del segmento. Los segmentos en donde se requieren más horas de mano de obra en pintura son una vez más F, TC y E, mientras que el que requiere menos horas es el segmento A.

Los valores oscilan entre las 0 horas hasta las 30-55 horas. La mediana tiene valores alrededor de las 4-5 horas, mientras que el 3º cuartil entre las 5,5-7,5 horas, lo que indica que en la mayoría de ocasiones, se requieren menos de 5,5-7,5 horas de mano de obra en pintura.Aquí también destaca la diferencia que hay entre el 3º cuartil y el valor máximo, en donde el máximo llega a ser más de 6 veces mayor. La desviación típica no es tan alta en esta ocasión, y oscila entre los 3-5,5 horas, lo que indica que estos datos están menos dispersos que el caso de los costes.

Horas de chapa

Horas Chapa	A	B	C	D	E	F	G	H	S	TA	TB	TC
Mínimo	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
1º cuartil	1,60	1,70	1,70	19,00	2,08	2,25	1,90	1,80	2,10	1,70	1,80	1,92
Mediana	2,91	3,10	3,08	3,20	3,44	3,70	3,50	3,30	3,50	3,10	3,10	3,40
Media	4,87	5,19	5,04	5,21	5,63	6,18	5,53	5,29	5,96	4,90	4,90	5,44
3º cuartil	6,00	6,46	6,25	6,40	6,60	7,30	7,12	6,70	6,80	6,00	5,90	6,60
Máximo	58,79	63,20	71,20	71,20	82,00	78,00	66,07	73,10	86,10	71,70	72,60	77,80
Desviación típica	5,70	6,03	6,44	5,87	6,81	7,19	6,01	5,93	7,64	5,78	5,86	6,31

Tabla 5: Estadísticos descriptivos de las horas de chapa de los siniestros.

En la tabla 5 se muestran los estadísticos para describir las horas de mano de obra en chapa. Las horas medias requeridas oscilan entre las 4,8-6 horas, dependiendo del segmento. Los segmentos en donde se requieren más horas de mano de obra en pintura son F y S , mientras que el que requiere menos horas es una vez más el segmento A.

El rango de horas se encuentra entre las 0 horas hasta las 58-86 horas. La mediana tiene valores alrededor de las 3-4 horas, mientras que el 3º cuartil entre las 6-7 horas, lo que indica

que en la mayoría de ocasiones, se requieren menos de 7 horas de mano de obra en pintura. Hay también una gran diferencia entre el 3º cuartil y el valor máximo, en donde el máximo llega a ser en ocasiones hasta 10 veces mayor. La desviación típica es bastante alta, lo que indica que hay una gran dispersión en estos datos.

4.2 Modelos predictivos con Deep Learning

A continuación se mostrarán los resultados obtenidos mediante los modelos de redes neuronales con una arquitectura Deep Learning para cada segmento de automóviles.

Antes de todo, es importante mencionar que se han probado múltiples configuraciones en las redes neuronales (número de capas, número de neuronas), con el objetivo de determinar con qué arquitectura se obtienen las mejores predicciones. No hay una regla general para ajustar los parámetros de la red (Le, 2015a), por lo que al final se eligió utilizar una arquitectura de tipo autoencoder con 4 capas ocultas: $(n, n/2, n/2, n)$, siendo n el número de variables de entrada, dado que fue el que mejores resultados obtuvo.

Es importante mencionar que en cada segmento, se ha probado proporcionar a la red neuronal como datos de entrada cada combinación posible de las operaciones de reparación de automóvil. Es decir, en los siniestros se disponen de 3 operaciones de reparación en las piezas: sustitución, pintura y reparación; se han creado diferentes modelos de redes neuronales, en las que cada modelo recibe como datos de entrada una combinación distinta de operaciones de reparación. Por lo tanto, se han obtenido 7 modelos según los regresores que tenga:

1. Red neuronal utilizando sustitución, pintura y reparación.
2. Red neuronal utilizando pintura y reparación.
3. Red neuronal utilizando pintura y sustitución.
4. Red neuronal utilizando reparación y sustitución.
5. Red neuronal utilizando sustitución.
6. Red neuronal utilizando pintura.
7. Red neuronal utilizando reparación.

A través de la creación de distintos modelos predictivos, se podrá obtener información bastante interesante de cada segmento, como es la de determinar qué modelo es el mejor predictor de las variables de interés, qué modelos no se pueden utilizar para predecir y qué operaciones de reparación aportan más información para la predicción.

Por otra parte, tal y como se ha visto en el estudio descriptivo de los datos del anterior apartado, aproximadamente los datos superiores al 3º cuartil indican casos de siniestros muy complicados que exigen unos costes y horas de trabajo altísimos. Por ello aparte de probar los modelos usando todos los datos, se ha estudiado también el subconjunto de datos que engloba desde el mínimo hasta el 3º cuartil.

Para simplificar la lectura, a la hora de hacer referencia a un modelo, ésta se indicará de la siguiente manera: [operación 1, operación 2, operación 3]. Por ejemplo, si se habla del modelo que usa todas las variables, se le referirá como modelo "[Pintura, Reparación,

Sustitución]"; si se habla de uno que solamente usa las variables de pintura y sustitución, será modelo "[Pintura, Sustitución]", y si se usa uno que solo usa la variable sustitución, se le hará referencia como modelo "[Sustitución]".

Por último, cabe mencionar que a la hora de referirnos a una operación (Sustitución, Reparación, Pintura), nos estamos refiriendo a las 27 zonas del automóvil en las que se realiza esa operación, por lo que por ejemplo la red neuronal [Sustitución], está recibiendo 27 atributos de entrada, referentes a las 27 zonas.

4.2.1 Coste total

A continuación se muestran los resultados obtenidos en los diferentes modelos utilizando todos los datos de cada segmento.

Error Coste Total	A	B	C	D	E	F	G
Pintura, Reparación, Sustitución	189,12 €	224,69 €	247,10 €	278,69 €	361,37 €	488,82 €	241,23 €
Pintura, Reparación	350,42 €	389,36 €	432,22 €	450,70 €	639,85 €	839,41 €	430,51 €
Pintura, Sustitución	186,92 €	222,93 €	255,82 €	265,73 €	339,81 €	476,07 €	238,90 €
Reparación, Sustitución	202,71 €	212,67 €	252,30 €	272,14 €	356,14 €	485,21 €	253,13 €
Sustitución	241,87 €	253,58 €	292,91 €	319,49 €	405,08 €	539,34 €	294,76 €
Pintura	402,64 €	430,62 €	469,77 €	521,83 €	691,78 €	864,34 €	493,29 €
Reparación	468,74 €	549,44 €	615,01 €	623,01 €	831,03 €	1.108,37 €	579,84 €

Error Coste Total	H	S	TA	TB	TC	PROMEDIO
Pintura, Reparación, Sustitución	307,49 €	353,78 €	279,17 €	325,04 €	367,52 €	305,33 €
Pintura, Reparación	494,33 €	733,85 €	532,24 €	568,13 €	708,93 €	547,50 €
Pintura, Sustitución	301,84 €	367,24 €	298,17 €	307,30 €	389,74 €	304,21 €
Reparación, Sustitución	316,25 €	381,94 €	338,11 €	339,32 €	398,36 €	317,36 €
Sustitución	371,95 €	409,79 €	347,80 €	377,52 €	438,22 €	357,69 €
Pintura	518,31 €	759,57 €	551,24 €	565,97 €	747,74 €	584,76 €
Reparación	603,65 €	885,35 €	665,22 €	679,63 €	861,38 €	705,89 €

Tabla 6: Error del coste total

Desviación típica error	A	B	C	D	E	F	G
Pintura, Reparación, Sustitución	273,53 €	310,51 €	350,28 €	371,86 €	540,07 €	743,39 €	355,31 €
Pintura, Reparación	519,54 €	644,81 €	777,14 €	763,62 €	1.082,33 €	1.386,47 €	694,71 €
Pintura, Sustitución	279,56 €	312,64 €	380,77 €	409,03 €	552,06 €	775,75 €	351,81 €
Reparación, Sustitución	268,34 €	301,09 €	357,54 €	381,76 €	547,83 €	781,36 €	354,99 €
Sustitución	281,68 €	296,23 €	370,55 €	388,10 €	549,94 €	798,68 €	363,74 €
Pintura	520,15 €	669,78 €	817,22 €	783,13 €	1.116,87 €	1.394,91 €	700,59 €
Reparación	641,99 €	759,94 €	896,12 €	907,68 €	1.283,83 €	1.651,15 €	813,01 €

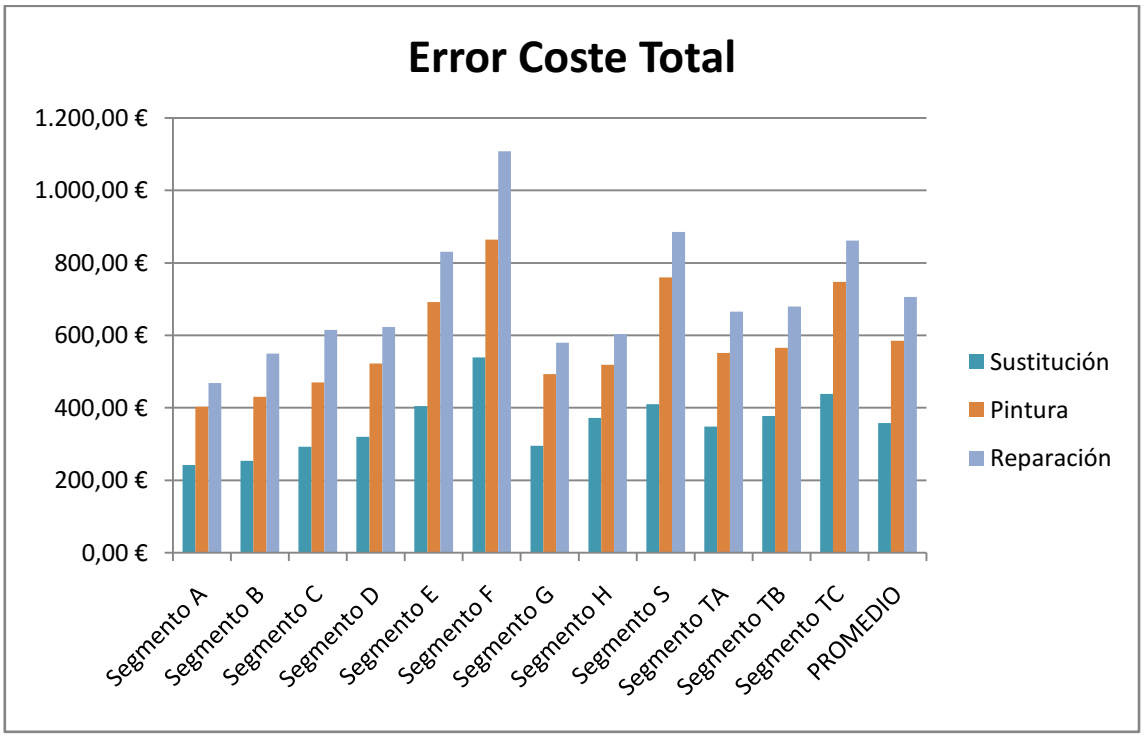
Desviación típica error	H	S	TA	TB	TC	PROMEDIO
Pintura, Reparación, Sustitución	451,97 €	451,97 €	443,70 €	479,64 €	598,00 €	447,52 €
Pintura, Reparación	809,60 €	809,60 €	839,51 €	948,65 €	1.194,02 €	872,50 €
Pintura, Sustitución	459,07 €	459,07 €	436,13 €	487,33 €	594,88 €	458,18 €
Reparación, Sustitución	461,13 €	461,13 €	416,86 €	488,04 €	603,01 €	451,92 €
Sustitución	477,70 €	477,70 €	427,83 €	501,01 €	615,22 €	462,37 €
Pintura	811,31 €	811,31 €	858,97 €	997,20 €	1.226,73 €	892,35 €
Reparación	897,87 €	897,87 €	997,51 €	1.120,17 €	1.377,67 €	1.020,40 €

Tabla 7: Desviación típica de los errores en el coste total

En la tabla 6 se muestran los errores medios que se han obtenido en cada segmento de automóviles. La primera columna indica los regresores utilizados en cada red neuronal. Las siguientes columnas hacen referencia a los doce segmentos de automóviles en los que se ha usado cada modelo y la última el error promedio de todos los segmentos. Los colores rojos hacen referencia a los peores modelos, mientras que los verdes a los mejores. Por otra parte, la tabla 7 muestra las desviaciones típicas de los errores obtenidos en la tabla 6.

Se puede observar en todos los segmentos, que las redes neuronales que usan un solo tipo de variable como datos de entrada, son los que más error tienen, junto a los que utilizan solamente pintura y reparación combinadas.

A continuación se muestra una gráfica en la que se comparan modelos que solo utilizan una variable:



Gráfica 1: Error del coste total en modelos individuales

Se puede observar tanto en la gráfica 1 como en la tabla 6 que [Reparación] es el peor modelo, tiene errores medios desde los 470€ hasta los 1110€, dependiendo del segmento del que hablemos, con un error de 705€ si hacemos la media en todos los segmentos. Seguido de [Reparación], el segundo modelo que peor predice es [Pintura], con errores medios entre los 400€ y los 900€, y con un error medio total de 584€. El modelo que mejor predice individualmente es por lo tanto, [Sustitución], cuyos errores rondan entre los 240€ hasta los 540€ en los segmentos y de media 357€, una cantidad considerablemente menor que en [Reparación] o [Pintura]. Con ello se concluye que la variable más importante en la reparación de automóviles es la sustitución de piezas. Es la variable que más información aporta a la hora de predecir el coste total. Esto quiere decir que la mayoría de veces, a la hora de reparar un automóvil, la opción más común es sustituir las piezas en la zona afectada. Por otra parte, la segunda variable más importante es la pintura, y por último, la reparación.

También se puede contemplar que el segmento F es donde se obtienen los mayores errores en la predicción, seguido por los segmento S, TC y E. En contraste, los segmentos A, B, C y G son donde se obtienen los errores más pequeños.

Volviendo a la Tabla 6, nos encontramos con diferencias particulares en cada segmento, aunque hay diversas características que se pueden agrupar. Se puede observar que en los segmentos C, S, TA, TB y TC, la red neuronal que mejor resultados ha conseguido es [Pintura, Reparación, Sustitución], es decir, la que ha recibido como datos de entrada todas las variables. Se han obtenido con ello errores medios de 247.1€, 353.78€, 278.17€, 325.04€ y 367.52€ respectivamente, lo que quiere decir que en media, la red neuronal se equivoca entre 250€-370€ por cada predicción.

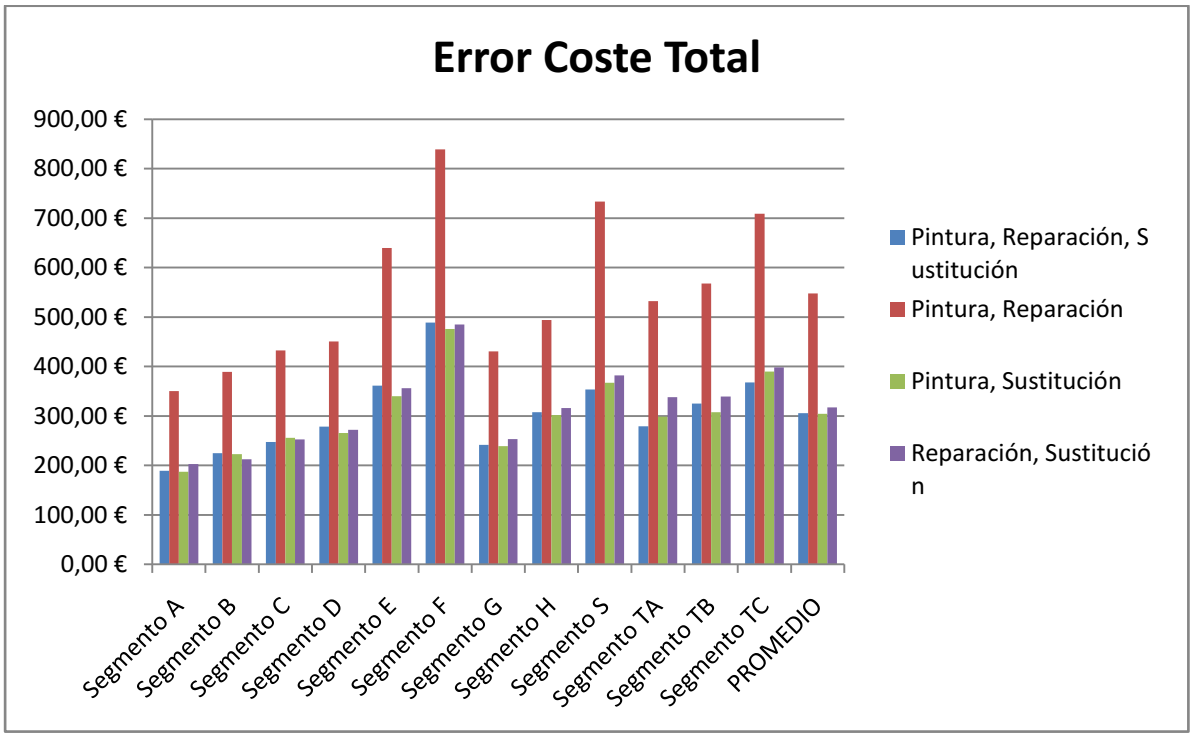
El segundo mejor modelo en estos segmentos es [Pintura, Sustitución], y ésta es seguida por [Reparación, Sustitución]; salvo en el segmento C, en donde [Reparación, Sustitución] es el que obtiene el segundo lugar (aunque la diferencia es mínima).

Otro grupo de segmentos en donde se pueden agrupar características similares, son los segmentos A, D, E, F, G y H. En todos estos segmentos, la mejor red neuronal es [Pintura, Sustitución], con errores medios de 186.92€, 265.73€, 339.81€, 476.07€, 238.90€ y 301.84€ respectivamente. Después, el segundo mejor modelo para los segmentos A, G, y H, es [Pintura, Reparación, Sustitución] y es seguido por poco por [Reparación, Sustitución]. Por otra parte, en los segmentos D, E, F [Reparación, Sustitución] es la segunda mejor, y es seguido por [Pintura, Reparación, Sustitución], aunque las diferencias tampoco son muy grandes.

Finalmente está el segmento B, que es el único segmento en donde las mejores predicciones se consiguen con el modelo [Reparación, Sustitución], con un error medio de 212,67€. Éste es seguido por [Pintura, Sustitución] y [Pintura, Reparación, Sustitución].

Por último, se puede observar que [Pintura, Reparación] es el que peor predice de todos en comparación con los demás modelos combinados. Éste llega a ser incluso peor que el modelo individual [Sustitución], lo que indica que solamente la variable sustitución, aporta más información e influye más en el coste final del siniestro que la pintura y la reparación juntas.

En la Gráfica 2, se puede comprobar que tal y como sucede con los modelos individuales, el segmento F es donde se obtienen los mayores errores de predicción, seguido por S, TC y E; y los segmentos A, B, C y G son los que se obtienen los errores más pequeños.



Gráfica 2: Error del coste total en modelos combinados

Por otra parte, mirando a tabla 7, se ve que las desviaciones típicas de los errores son bastante grandes y en todos los casos, mayores que los errores medios. Por ejemplo, en el caso del Segmento A, se tiene un error medio de 186,92€, mientras que la desviación típica del error es de 268,34€. Esto quiere decir que el error puede variar en unos 270€, por lo que no se sabrá si la predicción se está equivocando en 180€, o en 450€. [Pintura, Sustitución] ha conseguido un error medio de 304€, mientras que su desviación típica es de 458€, lo que hace que las predicciones sean bastante imprecisas.

Si ordenamos de menor a mayor los segmentos por su error de predicción, se obtiene la siguiente clasificación:

Segmento	Error
A	186,92 €
B	212,67 €
G	238,90 €
C	247,10 €
D	265,73 €
TA	279,17 €
H	301,84 €
TB	325,04 €
E	339,81 €
S	353,78 €
TC	367,52 €
F	476,07 €

Tabla 8: Ranking de segmentos por error absoluto en el coste total

Por otra parte, el ranking de los diferentes modelos de redes de neuronas según el error medio obtenido en todos los segmentos, queda de la siguiente manera:

Modelo	Error
Pintura, Sustitución	304,21 €
Pintura, Reparación, Sustitución	305,33 €
Reparación, Sustitución	317,36 €
Sustitución	357,69 €
Pintura, Reparación	547,50 €
Pintura	584,76 €
Reparación	705,89 €

Tabla 9: Ranking de modelos en el coste total por error absoluto

Esta tabla indica que si se tuviese que elegir solamente un modelo para tratar a todos los segmentos de automóviles, se usaría el modelo [Pintura, Sustitución], seguido de [Pintura, Reparación, Sustitución].

Análisis en términos relativos

Hasta ahora, se han mostrado los errores de las predicciones en términos monetarios. Pero para comprobar en qué medida éstas son buenas, es necesario comparar el error medio por segmento con el coste medio por segmento. Es decir, no es lo mismo equivocarse en 100€ para una reparación que ha costado10000€, que equivocarse en 100€ para una reparación que ha costado 200€. En el primero no estamos equivocando solamente en un 1%, mientras que en la segunda en un 50%. Por ello se muestra a continuación los errores y sus desviaciones típicas en términos relativos al coste medio de los segmentos.

Error Relativo Coste Total	A	B	C	D	E	F	G
Pintura, Reparación, Sustitución	73,59%	72,57%	72,62%	70,75%	68,41%	65,60%	74,30%
Pintura, Reparación	51,06%	52,46%	52,11%	52,70%	44,07%	40,93%	54,14%
Pintura, Sustitución	73,89%	72,78%	71,65%	72,11%	70,30%	66,50%	74,55%
Reparación, Sustitución	71,69%	74,03%	72,04%	71,44%	68,87%	65,85%	73,04%
Sustitución	66,22%	69,04%	67,54%	66,47%	64,59%	62,05%	68,60%
Pintura	43,76%	47,42%	47,95%	45,23%	39,53%	39,17%	47,46%
Reparación	34,53%	32,91%	31,85%	34,61%	27,36%	22,00%	38,24%

Error Coste Total	H	S	TA	TB	TC	PROMEDIO
Pintura, Reparación, Sustitución	68,68%	69,89%	69,56%	68,47%	71,20%	70,47%
Pintura, Reparación	49,65%	37,54%	41,96%	44,89%	44,44%	47,16%
Pintura, Sustitución	69,26%	68,75%	67,48%	70,19%	69,46%	70,58%
Reparación, Sustitución	67,79%	67,49%	63,13%	67,09%	68,78%	69,27%
Sustitución	62,12%	65,12%	62,07%	63,38%	65,66%	65,24%
Pintura	47,21%	35,36%	39,89%	45,11%	41,40%	43,29%
Reparación	38,52%	24,65%	27,46%	34,08%	32,49%	31,56%

Tabla 10: Precisión relativa del coste total

Desviación	A	B	C	D	E	F	G
Pintura, Reparación, Sustitución	38,20%	37,91%	42,77%	39,03%	47,21%	52,31%	37,85%
Pintura, Reparación	72,56%	78,73%	94,89%	80,14%	94,61%	97,57%	74,00%
Pintura, Sustitución	39,04%	38,17%	46,49%	42,93%	48,26%	54,59%	37,47%
Reparación, Sustitución	37,48%	36,76%	43,66%	40,07%	47,89%	54,99%	37,81%
Sustitución	39,34%	36,17%	45,24%	40,73%	48,07%	56,21%	38,75%
Pintura	72,65%	81,78%	99,78%	82,19%	97,63%	98,16%	74,63%
Reparación	89,66%	92,79%	109,42%	95,26%	112,22%	116,20%	86,60%

Desviación	H	S	TA	TB	TC	PROMEDIO
Pintura, Reparación, Sustitución	46,04%	38,47%	48,38%	46,52%	46,86%	43,46%
Pintura, Reparación	82,46%	68,90%	91,548%	92,01%	93,58%	85,08%
Pintura, Sustitución	46,76%	39,07%	47,56%	47,27%	46,62%	44,52%
Reparación, Sustitución	46,97%	39,25%	45,45%	47,34%	47,26%	43,74%
Sustitución	48,66%	40,66%	46,65%	48,59%	48,21%	44,77%
Pintura	82,64%	69,05%	93,67%	96,72%	96,14%	87,09%
Reparación	91,45%	76,41%	108,77%	108,65%	107,97%	99,62%

Tabla 11: Desviación relativa de los errores en el coste total

La tabla 10 muestra que en el mejor modelo ([Pintura, Sustitución]), las redes neuronales predicen con una exactitud media de entre un 65% a un 75%, dependiendo del segmento y de media entre todos los segmentos, un 70%.

Tal y como se ha visto antes, el modelo [Sustitución] es el que mejor resultados tiene individualmente . Solamente con él se consigue una precisión promedia del 65%. Por otra parte, se puede observar que el añadir las variables pintura y/o reparación al modelo [Sustitución], ya sean ambas o individualmente, se suele ganar solamente un 5% de precisión en las predicciones. Los modelos que no usan sustitución, se equivocan en más de un 50% del coste total del siniestro, por lo que no resultan de utilidad.

Respecto a las desviaciones típicas mostradas en la Tabla 11, éstas indican que los errores son bastante dispersos, y suelen oscilar aproximadamente en un 40% de la media del coste total, lo que indica que hay bastante incertidumbre en las predicciones. Una explicación a esta desviación típica tan alta, es el gran rango de costes con el que tiene que lidiar la red neuronal, tal y como se ha mostrado en la tabla 2 del apartado 4.1. Por último, se puede observar que para los modelos que predicen peor, las desviaciones típicas también son mayores, llegando en ocasiones a ser superiores que el coste medio del segmento.

Si ordenamos los segmentos por la precisión en sus predicciones, se obtiene la siguiente lista:

Segmento	Precisión
G	74,55%
B	74,03%
A	73,89%
C	72,62%
D	72,11%
TC	71,20%
E	70,30%
TB	70,19%
S	69,89%
TA	69,56%
H	69,26%
F	65,85%

Tabla 12: Ranking de segmentos por precisión en el coste total

Se puede ver que se obtiene un ranking totalmente distinto a la tabla 8 en el que solamente se tuvo en cuenta el error absoluto. El segmento G es donde mejor se predicen los resultados, con un error del 25%, seguido por los segmentos B, A, C, D, TC, E y TB, con errores menores del 30%. F queda en última posición, con una precisión solamente del 65%. Los segmentos S, TA y H obtienen precisiones prácticamente del 70%.

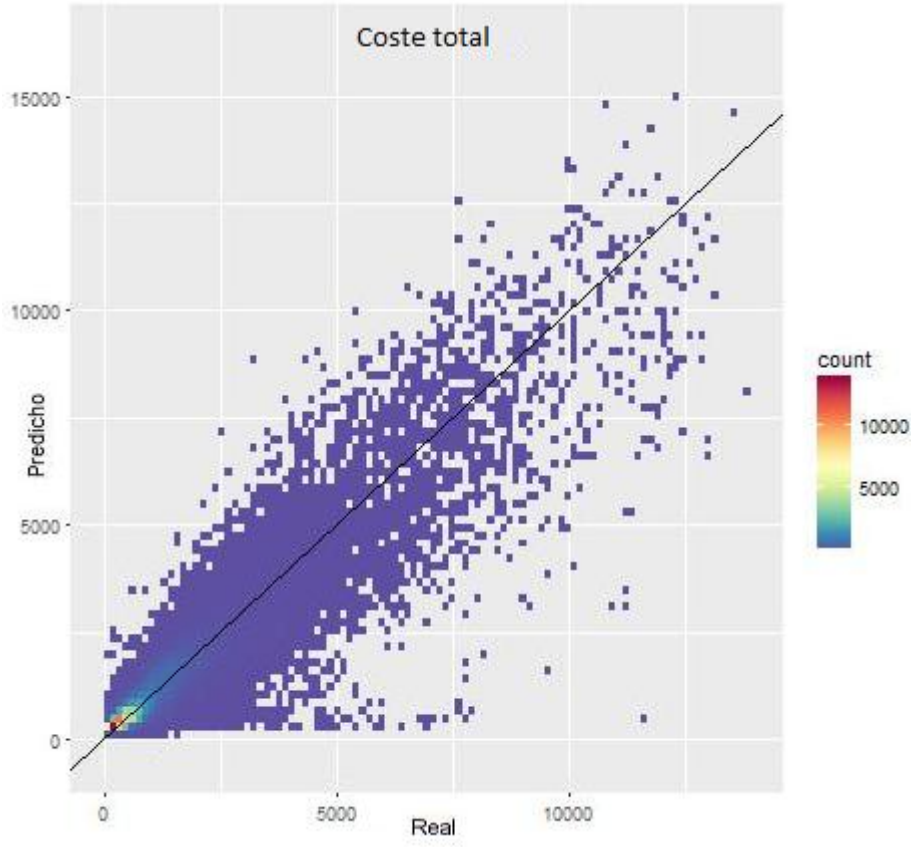
Por otra parte, el ranking de los diferentes modelos de redes de neuronas según la precisión obtenida entre todos los segmentos, queda de la siguiente manera:

Modelo	Precisión
Pintura, Sustitución	73,02%
Pintura, Reparación, Sustitución	72,75%
Reparación, Sustitución	72,04%
Sustitución	68,08%
Pintura, Reparación	51,52%
Pintura	47,80%
Reparación	37,00%

Tabla 13: Ranking de modelos por precisión en el coste total

Este ranking es el mismo que el obtenido usando errores absoluto de la tabla 9. El mejor modelo para predecir el coste es [Pintura, Sustitución], seguida muy cercanamente de [Pintura, Reparación, Sustitución].

A continuación se muestra una gráfica de las predicciones en el segmento G. En el eje X se pondrán los valores reales, mientras que en el eje Y, se pondrán los valores predichos por la red neuronal, de manera que si el valor predicho y y el valor real x son el mismo, entonces el punto (x, y) se encontrará en la recta $y = x$. Si el valor predicho es menor que el real, entonces el punto se situará por debajo de dicha recta, y si es mayor, se situará por encima.



Gráfica 3: Predicción y valor real en el coste total

El índice *count* situado en la leyenda de la Gráfica 3 indica cuántas veces se superponen los puntos en la gráfica. Esto quiere decir que los puntos rojos hay en realidad más de 10000

predicciones, los amarillos más de 5000, etc., mientras que los puntos morados indican una sola predicción. En esta gráfica se puede apreciar que hay una gran cantidad de puntos que se superponen para costes relativamente pequeños, en este caso, para costes menores de 1800€ aproximadamente. Este conjunto de puntos son los que más cercanos están en cuanto a la recta $y = x$, por lo que se considera que hay una gran cantidad de veces en el que el modelo está prediciendo bastante bien. Por otra parte, se puede observar que a medida que aumenta el valor del coste, hay más predicciones que se alejan de la recta. En este caso, esto quiere decir que la red neuronal, para valores pequeños predice bien, pero a medida que van aumentando los valores, sus predicciones empeoran.

Una causa que pueda explicar este patrón, es que tal y como se ha analizado en la tabla 2 del apartado 4.1, se ve que la mayoría de los siniestros tienen un coste relativamente bajo. En el caso del segmento G, el 3º cuartil del coste del siniestro tiene un valor de 1181€, lo que quiere decir que el 75% de los valores son menores de 1181€, mientras que el coste máximo es de 13940€.

Si analizamos el proceso aprendizaje de la red neuronal, ésta al recibir mayoritariamente instancias en los que el coste ha sido menor de 1200€, ha conseguido aprender a predecir correctamente estas instancias. Esto conlleva dos problemas: la primera es que no ha conseguido un buen aprendizaje para valores grandes, ya que la mayoría de veces aprende ejemplos de siniestros con costes pequeños, y la segunda es que el hecho de haber recibido información sobre siniestros excepcionales, ésta puede haber distorsionado sus predicciones para los valores pequeños. Por ello, se ha probado la ejecución de la mejor red neuronal para cada segmento, solamente en los datos desde el mínimo hasta el 3º cuartil..

Resultados en el subconjunto 0-3ºcuartil

Segmento	Coste medio	Error medio	Desviación típica	Precisión	Desviación relativa
A	369,00 €	84,20 €	75,37 €	77,18%	20,43%
B	414,10 €	92,14 €	83,19 €	77,75%	20,09%
C	445,90 €	105,20 €	97,12 €	76,41%	21,78%
D	476,50 €	117,20 €	106,55 €	75,40%	22,36%
E	533,70 €	143,73 €	128,28 €	73,07%	24,04%
F	654,00 €	195,98 €	173,13 €	70,03%	26,47%
G	494,40 €	117,45 €	107,97 €	76,24%	21,84%
H	511,20 €	137,32 €	116,01 €	73,14%	22,69%
S	529,10 €	142,33 €	134,84 €	73,10%	25,48%
TA	486,50 €	126,95 €	109,10 €	73,91%	22,43%
TB	507,40 €	141,86 €	123,97 €	72,04%	24,43%
TC	615,10 €	172,80 €	146,90 €	71,91%	23,88%
Promedio	503,08 €	131,43 €	116,87 €	74,18%	22,99%

Tabla 14: Resultados en el subconjunto 0-3º cuartil del coste total

Se puede observar que el coste medio por segmento es obviamente bastante menor, ya que se están cogiendo los datos hasta el 3º cuartil. El error medio se ha reducido a la más de la mitad: antes se obtenían errores desde los 180€ hasta los 500€ y con una media de 300€, mientras que ahora los errores rondan entre los 80€ y los 200€ y con una media de 131,43€

En la siguiente tabla se muestran las ganancias que se han obtenido en cada segmento al coger solamente este subconjunto de datos:

Segmento	0-3º cuartil		Todos los datos		Ganancia precisión	Ganancia desviación
	Precisión	Desviación típica	Precisión	Desviación típica		
A	77,18%	20,43%	73,89%	39%	3,29%	18,61%
B	77,75%	20,09%	74,03%	37%	3,72%	16,67%
C	76,41%	21,78%	72,62%	43%	3,79%	20,99%
D	75,40%	22,36%	72,11%	40%	3,29%	17,71%
E	73,07%	24,04%	70,30%	47%	2,77%	23,17%
F	70,03%	26,47%	65,85%	55%	4,18%	28,12%
G	76,24%	21,84%	74,55%	38%	1,69%	16,01%
H	73,14%	22,69%	69,26%	46%	3,88%	23,35%
S	73,10%	25,48%	69,89%	38%	3,21%	12,99%
TA	73,91%	22,43%	69,56%	48%	4,35%	25,96%
TB	72,04%	24,43%	70,19%	47%	1,85%	22,09%
TC	71,91%	23,88%	71,20%	47%	0,71%	22,98%
Promedio	74,18%	22,99%	71,12%	43,75%	3,06%	20,72%

Tabla 15: Ganancia en la predicción del coste total en el subconjunto 0-3º cuartil

Se puede observar que de media los modelos son un 3% más precisos que antes, salvo en los segmentos TB, TC y G, en donde la disminución del error relativo es mínima.

Pero el beneficio más importante no ha sido la tímida ganancia que se ha obtenido en la tasa de acierto, sino la gran disminución en la desviación típica del error. Mientras antes rondaba en desviaciones del 40% o más, ahora se encuentra en términos relativos entre el 20-25%, habiendo disminuido un 21% de media. Por ello se concluye que las predicciones en el subconjunto 0-3º cuartil son bastante más precisas que las predicciones usando todos los datos.

Finalmente, se muestra el ranking tanto en términos monetarios como en la precisión relativa.

Orden absoluto	Error absoluto	Orden relativo	Precisión
A	84,2	B	77,75%
B	92,14	A	77,18%
C	105,2	C	76,41%
D	117,20 €	G	76,24%
G	117,45 €	D	75,40%
TA	126,95	TA	73,91%
H	137,32	H	73,14%
TB	141,86	S	73,10%
S	142,33	E	73,07%
E	143,73	TB	72,04%
TC	172,8	TC	71,91%
F	195,98	F	70,03%

Tabla 16: Ranking de segmentos en el coste total en 0-3º cuartil

La clasificación por error y por precisión son muy parecidas. En términos monetarios, los segmentos que obtienen menor error son A, B y C, con una precisión del 77.18%, 77.75% y 76.41% respetivamente. Los segmentos en donde se predice peor son TC y F, con errores de 172 y 195€, y precisiones del 72 y 70%.

4.2.2 Coste de las piezas

En este apartado se mostrarán los resultados obtenidos en todos los segmentos de automóviles en la predicción del coste de las piezas de un siniestro.

Error Coste de Piezas	A	B	C	D	E	F	G
Pintura, Reparación, Sustitución	109,99 €	125,50 €	175,92 €	174,27 €	263,75 €	365,23 €	174,39 €
Pintura, Reparación	258,59 €	290,08 €	344,47 €	388,16 €	522,74 €	699,59 €	315,25 €
Pintura, Sustitución	116,57 €	129,47 €	162,23 €	165,62 €	266,88 €	360,37 €	166,23 €
Reparación, Sustitución	117,73 €	136,01 €	152,32 €	224,42 €	268,62 €	356,31 €	152,77 €
Sustitución	118,84 €	135,38 €	165,10 €	182,39 €	241,67 €	424,67 €	153,26 €
Pintura	246,64 €	318,98 €	367,99 €	391,19 €	506,79 €	735,77 €	327,39 €
Reparación	315,23 €	384,67 €	447,99 €	433,17 €	615,47 €	855,70 €	405,05 €

Error Coste de Piezas	G	H	S	TA	TB	TC	PROMEDIO
Pintura, Reparación, Sustitución	174,39 €	215,80 €	284,88 €	279,70 €	234,46 €	287,39 €	224,27 €
Pintura, Reparación	315,25 €	385,52 €	604,09 €	418,85 €	434,27 €	609,85 €	439,29 €
Pintura, Sustitución	166,23 €	205,49 €	241,77 €	209,23 €	244,69 €	336,37 €	217,08 €
Reparación, Sustitución	152,77 €	210,52 €	253,56 €	213,25 €	251,14 €	313,17 €	220,82 €
Sustitución	153,26 €	210,24 €	243,86 €	202,84 €	254,86 €	294,83 €	218,99 €
Pintura	327,39 €	400,85 €	559,43 €	451,88 €	456,96 €	595,71 €	446,63 €
Reparación	405,05 €	434,60 €	663,68 €	501,32 €	516,08 €	680,96 €	521,16 €

Tabla 17: Error absoluto del coste de piezas

Desviación típica error	A	B	C	D	E	F	G
Pintura, Reparación, Sustitución	201,24 €	237,91 €	295,43 €	297,91 €	438,56 €	640,85 €	280,35 €
Pintura, Reparación	401,31 €	518,51 €	634,65 €	599,64 €	881,03 €	1.170,92 €	554,97 €
Pintura, Sustitución	202,31 €	227,98 €	297,01 €	306,44 €	451,80 €	659,23 €	285,36 €
Reparación, Sustitución	202,02 €	236,06 €	285,66 €	306,29 €	448,21 €	658,55 €	280,05 €
Sustitución	203,41 €	245,16 €	302,15 €	314,34 €	480,15 €	690,54 €	283,94 €
Pintura	423,47 €	535,60 €	667,33 €	653,57 €	964,08 €	1.176,87 €	581,26 €
Reparación	471,28 €	582,84 €	724,62 €	738,56 €	1.061,73 €	1.408,66 €	633,85 €

Desviación típica error	H	S	TA	TB	TC	PROMEDIO
Pintura, Reparación, Sustitución	383,58 €	483,46 €	343,52 €	405,67 €	522,14 €	377,55 €
Pintura, Reparación	641,63 €	921,37 €	694,00 €	808,30 €	1.010,08 €	736,37 €
Pintura, Sustitución	387,31 €	515,00 €	366,52 €	420,26 €	498,08 €	384,78 €
Reparación, Sustitución	385,05 €	505,31 €	382,29 €	411,04 €	507,36 €	383,99 €
Sustitución	396,26 €	505,86 €	381,93 €	425,83 €	537,97 €	397,30 €
Pintura	664,31 €	992,95 €	712,73 €	830,63 €	1.067,25 €	772,50 €
Reparación	715,43 €	1.113,45 €	827,97 €	920,19 €	1.154,93 €	862,79 €

Tabla 18: Desviación típica de los errores en el coste de piezas

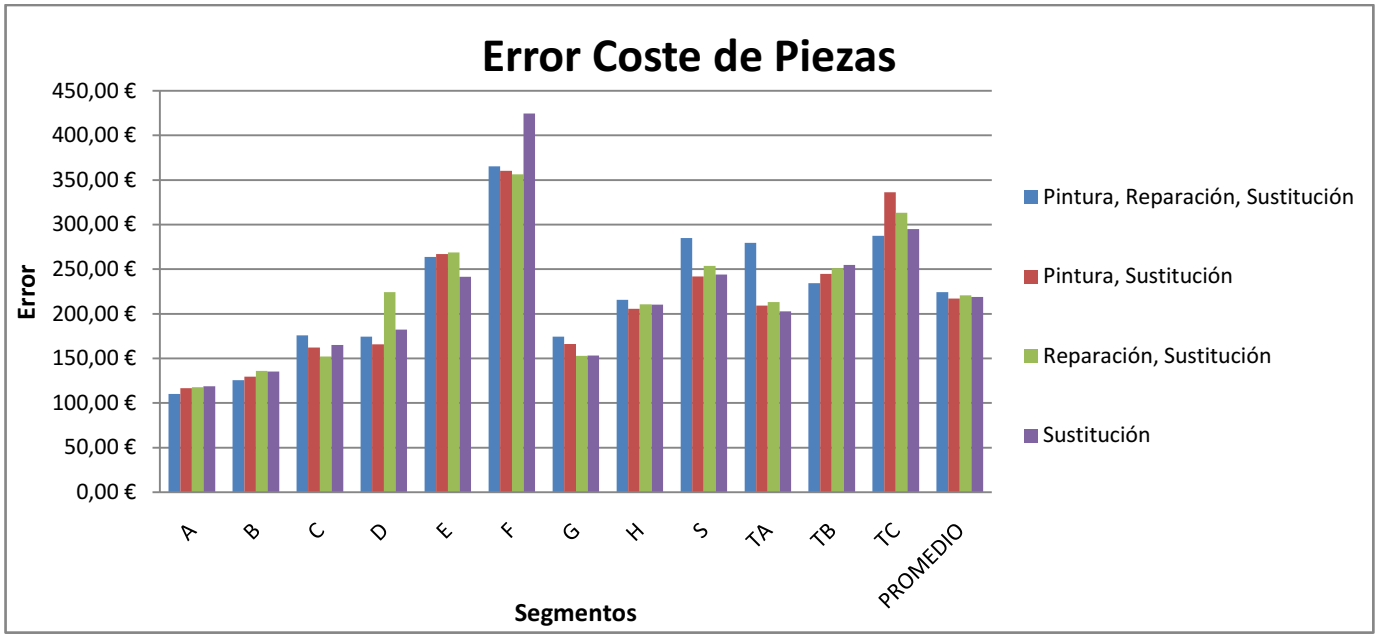
En las tablas 17 y 18, se muestran los errores medios que se han obtenido los diferentes modelos de redes neuronales para la predicción del coste de las piezas y sus desviaciones típicas respectivamente.

Se puede observar en todos los segmentos, tal y como sucede con las predicciones en el coste total, las redes neuronales que reciben solamente una variable como datos de entrada, son los que más error tienen en sus predicciones, con la excepción de [Sustitución], que en ocasiones llega a ser el mejor modelo. [Reparación] es el peor modelo que predice, tiene errores medios desde los 315€ hasta los 855€, dependiendo del segmento del que hablemos y con un error medio de 521€ si realizamos un promedio en todos los segmentos. Seguido de [Reparación], el segundo modelo que peor predice los resultados es [Pintura], con errores medios entre los 246€ y los 736€, y con un error medio total de 446,63€. El modelo que mejor predice individualmente es [Sustitución], cuyos errores rondan entre los 118€ hasta los 425€ y una media total de 218,99€, menos la mitad que la media en [Reparación] o [Pintura].

Tal y como sucede con las predicciones en el coste total, la variable que más información aporta a la hora de predecir el coste de las piezas, es la sustitución de piezas, lo cual es lógico, dado que el número de piezas sustituidas es directamente proporcional al coste de las mismas.

Por otra parte, se puede observar que los errores de [Pintura, Reparación] se encuentran entre los 260€ y los 610€, los cuales son sumamente peores que los obtenidos en [Sustitución]. Esto indica que la sustitución de piezas es más importante para predecir el coste de las piezas que la pintura y reparación combinadas.

A continuación, se muestran gráficamente los errores del coste de las piezas en cada segmento para los mejores modelos: [Pintura, Reparación Sustitución], [Pintura, Sustitución], [Reparación, Sustitución] y [Sustitución].



Gráfica 4: Mejores modelos en el coste de piezas

Se comprueba tanto en la gráfica 4 como en la tabla 17, que los modelos en el segmento F son donde se obtienen los mayores errores, seguido por los segmentos TC, S y E. En contraste, en los segmentos A, B, C y G son donde se obtienen los errores absolutos más pequeños.

En los segmentos A, B, TB y TC, la red neuronal que mejor resultados ha conseguido es [Pintura, Reparación, Sustitución]. Se han obtenido con ello errores medios de 109,99€, 125,50€, 234,46€ y 287,39€ respectivamente, lo que quiere decir que en media, esta red neuronal se equivoca entre 110-290€ por cada predicción. Para el segmento A, los restantes modelos ([Pintura, Sustitución], [Reparación, Sustitución] y [Sustitución]) tienen prácticamente el mismo error, con diferencias de 1-2€. En los segmentos B y TB, el segundo mejor modelo es [Pintura, Sustitución] y en TC es [Sustitución].

Respecto a los segmentos D, H y S, la mejor red neuronal es [Pintura, Sustitución], con errores medios de 165,62€, 205,49€ y 241,77€ respectivamente. El segundo mejor para el segmento D es [Pintura, Reparación,, Sustitución], y para H es [Reparación, Sustitución] junto a [Sustitución], y para S es [Sustitución].

Luego están los segmentos C, F y G, cuyo mejor modelo es [Reparación, Sustitución], con errores de 152,32€, 356,31€ y 152,77€ y por último están los segmentos E y TA, cuyo mejor modelo [Sustitución].

Respecto a las desviaciones típicas mostradas en la tabla 18, dependiendo del segmento éstas oscilan entre los 200-600€ en el mejor de los casos, y tiene un promedio entre segmentos de 377,55€ para el mejor modelo. Sus valores son bastante mayores que los errores medios, lo que indica que hay una gran dispersión en los mismos, por lo que las predicciones de estos modelos son muy imprecisas.

Si ordenamos de menor a mayor los segmentos según su error absoluto de predicción, se obtiene la siguiente lista:

Segmento	Error
A	109,99 €
B	125,50 €
C	152,32 €
G	152,77 €
D	165,62 €
TA	202,84 €
H	205,49 €
TB	234,46 €
E	241,67 €
S	241,77 €
TC	287,39 €
F	356,31 €

Tabla 19: Ranking de segmentos por error absoluto en el coste de piezas

Por otra parte, el ranking de los diferentes modelos de redes de neuronas según el error medio obtenido en todos los segmentos, queda de la siguiente manera.

Modelo	Error
Pintura, Sustitución	217,08 €
Sustitución	218,99 €
Reparación, Sustitución	220,82 €
Pintura, Reparación, Sustitución	224,27 €
Pintura, Reparación	439,29 €
Pintura	446,63 €
Reparación	521,16 €

Tabla 20: Ranking de modelos por error absoluto en el coste de piezas

En media, el mejor modelo es [Pintura, Sustitución], aunque tiene prácticamente el mismo error que [Sustitución].

Análisis en términos relativos

Como con los modelos que predicen el coste total, es interesante ver el error medio y sus desviaciones típicas en términos relativos:

Precisión	A	B	C	D	E	F	G
Pintura, Reparación, Sustitución	70,94%	71,37%	65,93%	67,38%	63,78%	62,01%	63,57%
Pintura, Reparación	31,68%	33,83%	33,29%	27,35%	28,22%	27,23%	34,15%
Pintura, Sustitución	69,20%	70,47%	68,58%	69,00%	63,35%	62,52%	65,27%
Reparación, Sustitución	68,89%	68,98%	70,50%	58,00%	63,11%	62,94%	68,09%
Sustitución	68,60%	69,12%	68,03%	65,86%	66,81%	55,83%	67,98%
Pintura	34,84%	27,24%	28,74%	26,79%	30,41%	23,47%	31,61%
Reparación	16,72%	12,26%	13,25%	18,93%	15,48%	10,99%	15,38%

Precisión	H	S	TA	TB	TC	PROMEDIO
Pintura, Reparación, Sustitución	60,14%	62,75%	53,63%	63,33%	64,98%	64,15%
Pintura, Reparación	28,79%	21,01%	30,56%	32,08%	25,69%	29,49%
Pintura, Sustitución	62,04%	68,39%	65,31%	61,73%	59,01%	65,41%
Reparación, Sustitución	61,12%	66,85%	64,65%	60,72%	61,84%	64,64%
Sustitución	61,17%	68,11%	66,37%	60,14%	64,08%	65,18%
Pintura	25,96%	26,85%	25,09%	28,53%	27,41%	28,08%
Reparación	19,73%	13,22%	16,89%	19,29%	17,03%	15,76%

Tabla 21: Error relativo del coste de piezas

Desviación	A	B	C	D	E	F	G
Pintura, Reparación, Sustitución	53,17%	54,27%	57,21%	55,76%	60,23%	66,66%	58,56%
Pintura, Reparación	106,03%	118,27%	122,90%	112,23%	120,99%	121,79%	115,93%
Pintura, Sustitución	53,45%	52,00%	57,52%	57,35%	62,04%	68,57%	59,61%
Reparación, Sustitución	53,37%	53,85%	55,32%	57,33%	61,55%	68,50%	58,50%
Sustitución	53,74%	55,92%	58,51%	58,83%	65,94%	71,83%	59,32%
Pintura	111,88%	122,17%	129,23%	122,32%	132,39%	122,41%	121,42%
Reparación	124,51%	132,95%	140,32%	138,23%	145,80%	146,52%	132,41%

Desviación	H	S	TA	TB	TC	PROMEDIO
Pintura, Reparación, Sustitución	70,85%	63,21%	56,95%	63,45%	63,62%	60,33%
Pintura, Reparación	118,51%	120,47%	115,05%	126,42%	123,08%	118,47%
Pintura, Sustitución	71,54%	67,34%	60,76%	65,73%	60,69%	61,38%
Reparación, Sustitución	71,12%	66,07%	63,38%	64,28%	61,82%	61,26%
Sustitución	73,19%	66,14%	63,32%	66,60%	65,55%	63,24%
Pintura	122,70%	129,83%	118,16%	129,91%	130,04%	124,37%
Reparación	132,15%	145,59%	137,26%	143,92%	140,72%	138,36%

Tabla 22: Desviación relativa de los errores en el coste de piezas

En la tabla 21 se puede observar que en el mejor de los casos, las redes neuronales predicen con una exactitud de entre un 62% a un 72%, dependiendo del segmento, y el mejor modelo tiene una precisión media del 65,41%.

El modelo [Sustitución] es el que mejor resultados tiene individualmente consiguiendo una precisión promedia del 65,18%. Los modelos que no usan sustitución, se equivocan en más de un 70% del coste medio de piezas, por lo que no resultan de utilidad. Además, se puede observar que el añadir la reparación o la pintura al modelo [Sustitución], apenas da ganancias, y de hecho en ocasiones, empeoran los resultados.

En cuanto a la desviación típica mostrada en la tabla 19, en los mejores modelos estás llegan a valores entre el 50% y el 70% de la media del coste de piezas, y con una media del 60% del coste. Esto quiere decir que predicciones en estos datos llevan consigo demasiada incertidumbre. La explicación a esta desviación típica tan alta, es la misma que en el coste total: existe un gran rango de costes con el que tiene que lidiar la red neuronal. lo que empeora la predicción en los modelos. Por otra parte, se ve que los modelos que no utilizan la sustitución, tienen desviaciones típicas mayores del 100%, es decir, que son incluso mayores que el coste medio de piezas

En la siguiente tabla se muestra el ranking de modelos en la predicción del coste de pieza, el cual es el mismo que el obtenido al comparar los modelos en términos monetarios. Se ve que el mejor modelo para predecir el coste de las piezas es [Pintura, ´Sustitución].

Modelo	Precisión
Pintura, Sustitución	65,41%
Sustitución	65,18%
Reparación, Sustitución	64,64%
Pintura, Reparación, Sustitución	64,15%
Pintura, Reparación	29,49%
Pintura	28,08%
Reparación	15,76%

Tabla 23: Ranking de modelos en el coste de piezas

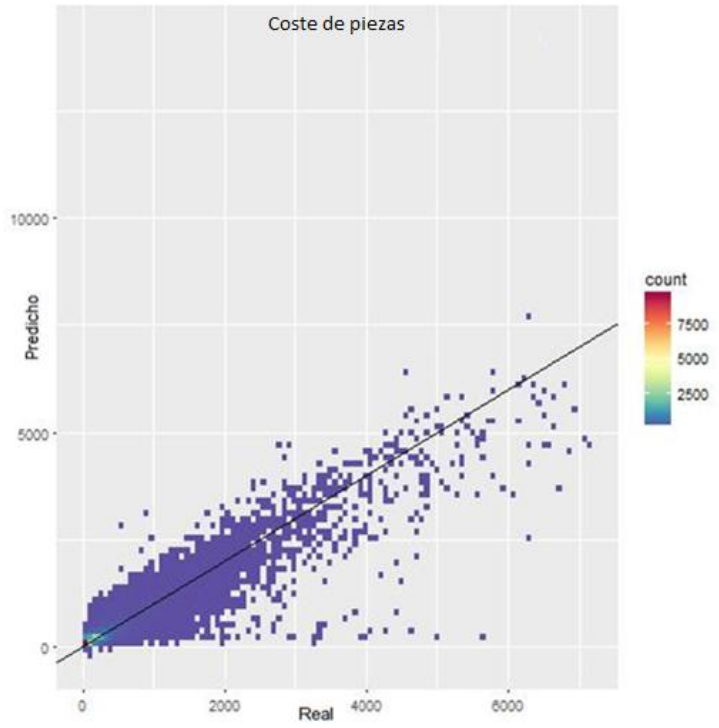
Si ordenamos los segmentos por la precisión en sus predicciones, se obtiene la siguiente lista:

Segmento	Precisión
A	70,94%
B	71,37%
C	70,50%
D	69,00%
G	68,09%
TA	66,37%
S	68,39%
E	66,81%
TC	64,98%
TB	63,33%
H	62,04%
F	62,94%

Tabla 24: Ranking de segmentos por precisión en el coste de piezas

Los segmentos A, B y C son donde se obtienen los mejores resultados, mientras que los peores se obtienen en H y F.

A continuación se muestra una gráfica de las predicciones en el segmento A. Es el mismo tipo que el mostrado en el gráfico 3 en donde se muestran los puntos (valor real, valor predicho), por lo que si el valor predicho y el valor real son el mismo, se situarán en la recta $y = x$, y cuanto más alejados de la recta estén, peor es la predicción.



Gráfica 5: Predicción y valor real en el coste de piezas en el segmento A

Aquí nos encontramos con una situación similar a la predicción del coste total. Hay un conjunto de puntos para los valores predichos más pequeños que se concentran en los valores más bajos de la gráfica y están muy cercanos a la recta $y = x$. También destaca el punto rojo que se encuentra prácticamente en el punto (0,0). Esto quiere decir que en una gran cantidad de ocasiones en el que el coste de piezas es sido 0.

Como se ha mostrado en la tabla 3 de los estadísticos descriptivos del coste de piezas, el 3º cuartil del segmento A es de 389€, el cual coincide aproximadamente conjunto de puntos verdes y amarillos. Por lo tanto, como en el coste total, se ha probado a ejecutar la mejor red neuronal de cada segmento, solamente para los datos hasta el tercer cuartil.

Resultados en el subconjunto mínimo-3ºcuartil

Segmento	Coste medio	Error medio	Desviación típica	Precisión	Desviación relativa
A	156,00 €	43,19 €	60,11 €	72,31%	38,53%
B	165,10 €	45,36 €	64,92 €	72,52%	39,32%
C	180,20 €	53,11 €	65,15 €	70,53%	36,16%
D	213,44 €	65,21 €	71,23 €	69,45%	33,37%
E	251,92 €	84,43 €	94,63 €	66,48%	37,56%
F	315,50 €	117,08 €	130,09 €	62,89%	41,23%
G	197,86 €	62,24 €	63,70 €	68,54%	32,20%
H	215,93 €	80,28 €	79,40 €	62,82%	36,77%
S	237,98 €	73,14 €	104,70 €	69,27%	44,00%
TA	234,93 €	77,67 €	82,11 €	66,94%	34,95%
TB	278,18 €	92,25 €	92,08 €	66,84%	33,10%
TC	303,36 €	103,92 €	116,44 €	65,74%	38,38%
Promedio	229,20 €	74,82 €	85,38 €	0,68 €	0,37 €

Tabla 25: Resultados del coste de piezas en el 0-3º cuartil

Se ve que coste medio de piezas es bastante menor, ya que se están cogiendo los datos hasta el 3º cuartil. El error medio en términos monetarios ha disminuido a casi un tercio: antes se obtenían errores desde los 110€ hasta los 360€ y una media de 206€, mientras que ahora los errores se encuentran entre los 40€ y los 118€ y con una media de 74,82€. Ocurre lo mismo con la desviación típica: antes oscilaba entre los 200 y los 500€, mientras que ahora se ha reducido al rango de 60-130€ y un promedio de 85,38€. Pero aunque ésta se haya reducido de esta manera, la desviación típica sigue siendo mayor que el error medio obtenido, por lo que las predicciones en este subconjunto de datos siguen llevando consigo bastante incertidumbre.

En la siguiente tabla se muestran las ganancias que se han obtenido en cada segmento al coger solamente este subconjunto de datos.

Segmento	0-3º cuartil		Todos los datos		Ganancia precisión	Ganancia desviación
	Precisión	Desviación típica	Precisión	Desviación típica		
A	72,31%	38,53%	70,94%	53,17%	1,37%	14,64%
B	72,52%	39,32%	71,37%	52,00%	1,15%	12,68%
C	70,53%	36,16%	70,50%	55,32%	0,03%	19,16%
D	69,45%	33,37%	69,00%	55,76%	0,45%	22,39%
E	66,48%	37,56%	66,81%	65,94%	-0,33%	28,38%
F	62,79%	41,23%	62,94%	68,50%	-0,05%	27,27%
G	68,54%	32,20%	68,09%	58,50%	0,45%	26,30%
H	62,82%	36,77%	62,04%	71,54%	0,78%	34,77%
S	69,27%	44,00%	68,39%	67,34%	0,88%	23,34%
TA	66,94%	34,95%	66,37%	63,32%	0,57%	28,37%
TB	66,84%	33,10%	66,33%	63,45%	0,51%	30,35%
TC	65,74%	38,38%	64,98%	63,62%	0,76%	25,24%
Promedio	67,85%	37,13%	67,31%	61,54%	0,55%	24,41%

Tabla 26: Ganancia por segmentar en subconjuntos en el coste de piezas

Se puede contemplar que no hay ganancias significativas en términos de precisión, pero la desviación típica ha disminuido entre un 15-35%. Pero estas desviaciones rondan en valores en torno al 40% del coste medio de piezas, lo cual sigue siendo un valor bastante alto.

Por último, se muestra el ranking por segmentos en las predicciones de este subconjunto de datos:

Orden absoluto	Error absoluto	Orden relativo	Precisión
A	43,19€	C	72,53%
B	45,36€	B	72,52%
C	53,11€	A	70,31%
G	62,24€	D	69,45%
D	65,21€	S	69,27%
S	73,14€	G	68,54%
TA	77,67€	TA	66,94%
H	80,28€	TB	66,84%
E	84,43€	TC	66,74%
TB	92,25€	E	66,48%
TC	103,92€	H	62,82%
F	117,08€	F	62,79%

Tabla 27: Ranking de segmentos en el coste de piezas en el subconjunto 0-3º cuaritl

Los segmentos en donde se obtienen las mejores predicciones siguen siendo A, B y C, son errores de 43.19€, 45.36€ y 53.11€ y precisiones del 72.31%, 72.52% y 72,53%

respectivamente. Por otra parte, el segmento en donde se obtienen los peores resultados sigue siendo F, con un error de 117,08€ y una precisión del 62,79%.

4.2.3 Horas de pintura

En este apartado analizarán los resultados obtenidos en la predicción de las horas de mano de obra en pintura en los diferentes segmentos de automóviles.

Error Horas Pintura	A	B	C	D	E	F	G
Pintura, Reparación, Sustitución	0,86	0,87	0,99	0,94	1,15	1,50	1,01
Pintura, Reparación	0,89	0,91	1,01	0,95	1,16	1,52	1,05
Pintura, Sustitución	0,88	0,89	0,98	1,00	1,19	1,47	1,07
Reparación, Sustitución	1,28	1,40	1,57	1,65	1,87	2,21	1,62
Sustitución	1,81	2,04	2,16	2,31	2,62	2,92	2,24
Pintura	0,97	1,05	1,06	1,10	1,23	1,58	1,16
Reparación	1,63	1,91	2,00	2,27	2,46	2,71	2,14

Error Horas Pintura	H	S	TA	TB	TC	PROMEDIO
Pintura, Reparación, Sustitución	1,27	1,17	1,01	1,19	1,20	1,10
Pintura, Reparación	1,30	1,21	1,02	1,16	1,24	1,12
Pintura, Sustitución	1,25	1,19	1,04	1,22	1,24	1,12
Reparación, Sustitución	1,80	1,87	1,67	1,89	1,99	1,73
Sustitución	2,66	2,56	2,34	2,63	2,68	2,41
Pintura	1,40	1,27	1,21	1,27	1,23	1,21
Reparación	2,26	2,28	2,09	2,36	2,50	2,22

Tabla 28: Error absoluto en las horas de pintura

Desviación Típica del Error	A	B	C	D	E	F	G
Pintura, Reparación, Sustitución	1,15	1,24	1,39	1,49	1,83	2,12	1,51
Pintura, Reparación	1,19	1,34	1,52	1,56	1,88	2,19	1,64
Pintura, Sustitución	1,23	1,34	1,49	1,55	1,86	2,20	1,58
Reparación, Sustitución	1,53	1,81	2,09	2,17	2,50	2,86	2,09
Sustitución	1,99	2,18	2,43	2,60	2,89	3,14	2,48
Pintura	1,31	1,50	1,64	1,62	1,93	2,27	1,69
Reparación	1,70	2,13	2,48	2,60	2,80	3,13	2,47

Desviación Típica del Error	H	S	TA	TB	TC	PROMEDIO
Pintura, Reparación, Sustitución	1,76	1,85	1,44	1,70	1,88	1,62
Pintura, Reparación	1,79	1,86	1,50	1,75	1,87	1,67
Pintura, Sustitución	1,84	1,84	1,48	1,75	1,91	1,67
Reparación, Sustitución	2,33	2,35	2,05	2,37	2,54	2,23
Sustitución	2,82	2,68	2,60	2,89	2,91	2,64
Pintura	1,97	1,86	1,67	1,79	1,95	1,77
Reparación	2,65	2,77	2,42	2,74	2,85	2,56

Tabla 29: Desviación típica de los errores en las horas de pintura

En las tablas 28 y 29, se muestran los errores medios y sus desviaciones típicas que se han obtenido en los modelos de redes neuronales para la predicción de las horas de pintura.

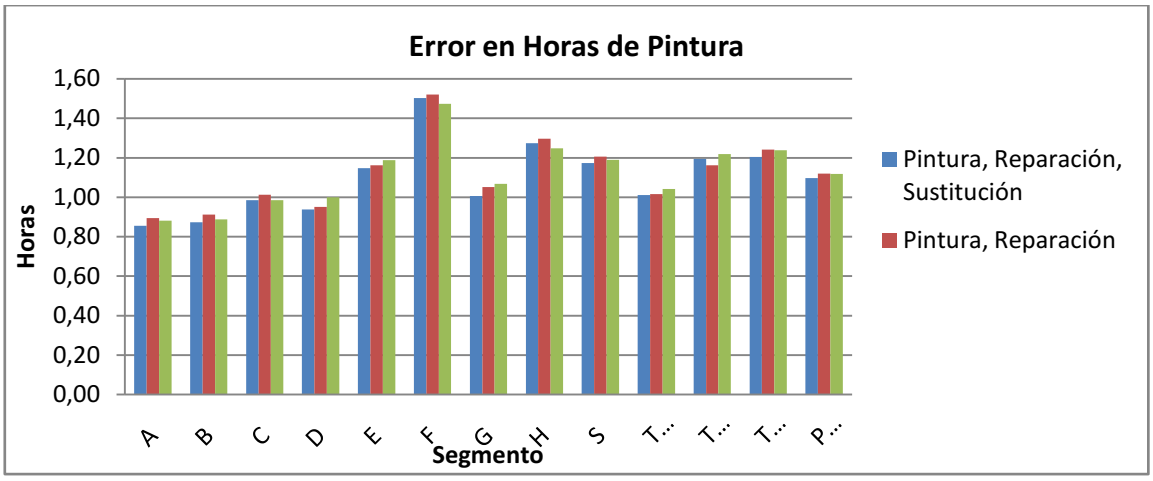
Tal y como sucede en los casos anteriores, las redes neuronales que reciben solamente una variable como datos de entrada, son los que mayor error tienen en sus predicciones, pero en esta ocasión la excepción es [Pintura]. [Sustitución] es el peor modelo que predice: tiene errores medios desde las 1,8 horas hasta las 2,7 horas, dependiendo del segmento del que hablemos, y un error medio de 2,4 horas. Seguido de [Sustitución], el segundo modelo que peor predice es [Reparación], con errores medios entre las 1,6 horas y las 2,7 horas y con un error medio total de 2,22 horas.

El modelo que mejor predice mejor individualmente es [Pintura], cuyos errores rondan entre la hora y la hora y media y con un error medio en todos los segmentos de 1,2 horas. Lógicamente la pintura de piezas es la variable que más información aporta dado que el número de piezas pintadas es directamente proporcional a las horas requeridas en pintarlas. El que menos aporta para predecir esta variable es la sustitución de piezas. Esta característica también tiene una explicación lógica: cuando se sustituye una pieza generalmente no es necesario pintarlo, salvo que la sustitución sea de una chapa y ésta tenga un color diferente al del automóvil.

Por otra parte, se puede observar que los errores de [Reparación, Sustitución] se encuentran entre las 1,3 horas y las 2,2 horas, los cuales son sumamente peores que los obtenidos en [Pintura]. Esto demuestra que la pintura de piezas es más importante para predecir las horas de pintura que la sustitución y reparación combinadas.

Por último, cabe mencionar que los modelos que no usan la variable pintura son los que peor predicen. De todas formas, el añadir de sustitución y/o reparación al modelo [Pintura], aporta mejorías de 0,1 horas en la predicción, por lo que no viene mal incluirlas.

A continuación, se muestran gráficamente los errores de las horas de pintura para los modelos con menor error: [Pintura, Reparación Sustitución], [Pintura, Sustitución], y [Pintura, Reparación].



Gráfica 6: Mejores modelos para Horas de Pintura

Se puede observar tanto en la gráfica 6 como en la tabla 28, que los modelos en el segmento F son donde se obtienen los mayores errores, seguido por los segmentos H, TC, S, TB y E. Por otra parte, en los segmentos A, B, C y G son donde se obtienen los menores errores.

El mejor modelo en la mayoría de los segmentos es [Pintura, Reparación, Sustitución], en donde se obtienen errores medios de entre 0.86 horas a 1.2 horas. Además, con este modelo se obtiene un error medio entre todos los segmentos de 1,1 horas.

Respecto a los segmentos C, F y H, la mejor red neuronal es [Pintura, Sustitución], con errores medios de 0.98, 1.47 y 1.25 horas respectivamente. Esta red neuronal obtiene un error promedio de 1.12 horas en total, casi lo mismo que el anterior modelo.

Por último, el segmento TB es el único en donde [Pintura, Reparación] tiene la mejor predicción. Este modelo también tiene un promedio global de 1.12 horas.

En cuanto a las desviaciones típicas de la tabla 29, se puede contemplar que éstos suelen ser mayor que los errores medios de predicción. Obtienen valores entre 1 y 2 horas, mientras que el error medio obtenido es igual o menor que 1,2 horas, por lo que aumenta de manera considerable la incertidumbre en las predicciones.

Si ordenamos de menor a mayor los segmentos según las horas erróneas de predicción, se obtiene la siguiente clasificación.

Segmento	Error
A	0,86
B	0,87
D	0,94
C	0,98
G	1,01
TA	1,01
E	1,15
TB	1,16
S	1,17
TC	1,2
H	1,25
F	1,47

Tabla 30: Ranking de segmentos por horas en las horas de pintura

A continuación se muestra el ranking de los diferentes modelos de redes de neuronas según el error medio obtenido en todos los segmentos. Como se puede ver, para predecir las horas de pintura el mejor modelo a utilizar es [Pintura, Reparación, Sustitución].

Modelo	Error
Pintura, Reparación, Sustitución	1,10
Pintura, Sustitución	1,118
Pintura, Reparación	1,119
Pintura	1,21
Reparación, Sustitución	1,73
Reparación	2,22
Sustitución	2,41

Tabla 31: Ranking de modelos por error absoluto en horas de pintura

Análisis en términos relativos

A continuación se muestran los resultados en términos de precisión para comparar los errores con las horas de pintura medias que se requiere normalmente en la reparación del automóvil tras el siniestro.

Error Horas Pintura	A	B	C	D	E	F	G
Pintura, Reparación, Sustitución	81,33%	83,52%	82,48%	84,55%	81,71%	77,38%	83,47%
Pintura, Reparación	80,48%	82,79%	82,00%	84,32%	81,46%	77,09%	82,71%
Pintura, Sustitución	80,78%	83,24%	82,49%	83,52%	81,06%	77,82%	82,46%
Reparación, Sustitución	72,13%	73,51%	72,03%	72,88%	70,22%	66,74%	73,43%
Sustitución	60,59%	61,57%	61,69%	61,90%	58,16%	55,99%	63,22%
Pintura	78,83%	80,24%	81,14%	81,94%	80,34%	76,21%	80,97%
Reparación	64,44%	64,03%	64,42%	62,52%	60,81%	59,14%	64,81%

Error Horas Pintura	H	S	TA	TB	TC	PROMEDIO
Pintura, Reparación, Sustitución	80,40%	80,03%	82,29%	80,06%	80,89%	81,51%
Pintura, Reparación	80,03%	79,46%	82,20%	80,61%	80,30%	81,12%
Pintura, Sustitución	80,80%	79,74%	81,76%	79,66%	80,33%	81,14%
Reparación, Sustitución	72,37%	68,17%	70,68%	68,44%	68,42%	70,75%
Sustitución	59,07%	56,39%	59,05%	56,03%	57,41%	59,26%
Pintura	78,49%	78,30%	78,82%	78,85%	80,50%	79,55%
Reparación	65,14%	61,09%	63,32%	60,62%	60,36%	62,56%

Tabla 32: Error relativo en horas de pintura

Desviación Típica del Error	A	B	C	D	E	F	G
Pintura, Reparación, Sustitución	25,20%	23,45%	24,79%	24,57%	29,22%	31,96%	24,85%
Pintura, Reparación	25,89%	25,23%	27,00%	25,68%	30,05%	33,03%	27,00%
Pintura, Sustitución	26,89%	25,31%	26,46%	25,57%	29,67%	33,19%	26,04%
Reparación, Sustitución	33,47%	34,15%	37,22%	35,76%	39,84%	43,03%	34,41%
Sustitución	43,46%	41,08%	43,24%	42,77%	46,13%	47,35%	40,75%
Pintura	28,53%	28,37%	29,12%	26,66%	30,83%	34,14%	27,83%
Reparación	37,04%	40,20%	44,13%	42,85%	44,58%	47,14%	40,66%

Desviación Típica del Error	H	S	TA	TB	TC	PROMEDIO
Pintura, Reparación, Sustitución	27,05%	31,57%	25,21%	28,38%	29,89%	27,18%
Pintura, Reparación	27,54%	31,72%	26,20%	29,14%	29,71%	28,18%
Pintura, Sustitución	28,38%	31,33%	25,90%	29,24%	30,33%	28,19%
Reparación, Sustitución	35,94%	40,07%	35,85%	39,53%	40,38%	37,47%
Sustitución	43,38%	45,72%	45,60%	48,26%	46,28%	44,50%
Pintura	30,27%	31,67%	29,31%	29,92%	30,93%	29,80%
Reparación	40,79%	47,20%	42,44%	45,75%	45,27%	43,17%

Tabla 33: Desviación típica relativa de los errores en las horas de pintura

En la tabla 32 se puede observar que el modelo [Pintura] es el que mejor resultados tiene individualmente consiguiendo una precisión media entre todos los segmentos del 79,55%. Por otra parte, cabe mencionar que el añadir las variables reparación y/o sustitución a [Pintura], se obtienen ganancias limitadas a un 2-3%

En los mejores modelos para cada segmento, las redes neuronales predicen con precisiones entre un 77 al 85% y, en el mejor modelo ([Pintura, Reparación, Sustitución]), se obtiene una precisión promedia del 81,51%.

A continuación se muestra la clasificación de los segmentos según la mejor precisión obtenida en cada segmento. En todos los segmentos se han obtenido precisiones mayores del 80%, salvo en F que se obtiene el 77,82%.

Segmento	Precisión
D	84,55%
B	83,52%
G	83,47%
C	82,49%
TA	82,39%
E	81,71%
A	81,33%
TC	80,89%
H	80,80%
TB	80,61%
S	80,03%
F	77,82%

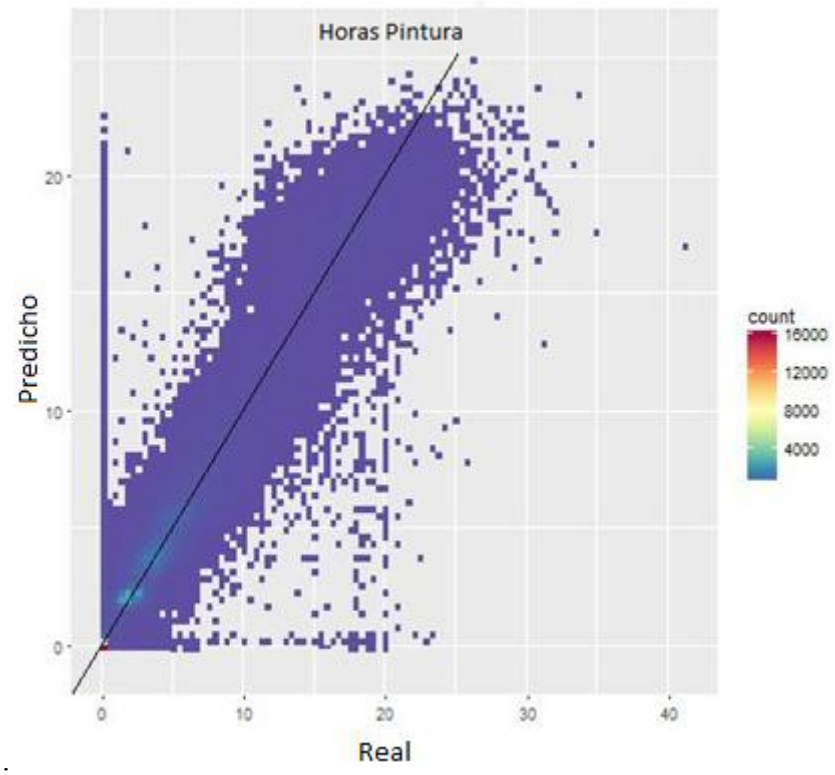
Tabla 34: Ranking de segmentos por precisión en las horas de pintura

Por otra parte, el ranking de modelos por error relativo queda de la siguiente manera:

Modelo	Error
Pintura, Reparación, Sustitución	81,51%
Pintura, Sustitución	81,14%
Pintura, Reparación	81,12%
Pintura	79,55%
Reparación, Sustitución	70,75%
Reparación	62,56%
Sustitución	59,26%

Tabla 35: Ranking de modelos en las horas de pintura por error relativo

A continuación se mostrará una gráfica de las predicciones en el segmento D. Se trata del mismo tipo de gráfico que los mostrados en los apartados anteriores, en donde se muestran los puntos (valor predicho, valor real), por lo que si el valor predicho y el valor real son el mismo, se situarán en la recta $y = x$, y cuanto más alejados de la recta estén, peor ha sido la predicción.



Gráfica 7: Horas de pintura: Valor real y valor predicho en el segmento D

En esta gráfica destaca el punto rojo que se sitúa en el (0,0), que indica que unos 16.000 siniestros en los que no se ha requerido horas de mano de obra en pintura. También destaca la linea vertical que se encuentra en el eje y, en donde en ocasiones se ha predicho que se requería un número determinado de horas en pintura, pero en realidad no se ha necesitado ninguna. Lo mismo sucede con los puntos en el eje de abscisas, en donde se ha predicho 0

horas, pero en realidad sí que se requería mano de obra en pintura. Se observa por otra parte que a medida que se tiene que predecir siniestros que requieren mayor trabajo en pintura, los puntos se alejan más de la recta, lo que quiere decir que predice peor.

Por otra parte, se ve una cantidad considerable de puntos que se amontonan y se encuentran muy cerca de la recta (los puntos de azul claro). Éstos son los valores menores de 6-7 horas.

Resultados en el subconjunto mínimo-3ºcuartil

Segmento	Tiempo medio	Error medio	Desviación típica	Precisión	Desviación relativa
A	3,15	0,58	0,48	81,52%	15,36%
B	3,52	0,58	0,48	83,51%	13,59%
C	3,61	0,62	0,53	82,78%	14,79%
D	3,86	0,64	0,55	83,53%	14,20%
E	3,87	0,71	0,6	81,62%	15,59%
F	3,6	0,81	0,68	77,56%	18,82%
G	3,71	0,61	0,55	83,56%	14,95%
H	3,82	0,72	0,65	81,15%	16,89%
S	3,48	0,68	0,58	80,48%	16,66%
TA	3,44	0,65	0,57	81,10%	16,65%
TB	3,83	0,74	0,63	80,64%	16,58%
TC	3,98	0,74	0,62	81,51%	15,69%
Promedio	3,66	0,67	0,58	81,58%	15,81%

Tabla 36: Resultados de horas de pintura en el subconjunto 0-3º cuartil

Se puede contemplar una considerable disminución en el error medio: ahora se encuentra en valores entre las 0,6-0,8 horas, con una media de 0,67, mientras que antes se encontraba entre las 0,86-1,5 horas. La desviación típica se ha reducido aproximadamente a la mitad: antes en los mejores modelos se encontraba en las 1-2 horas, mientras que ahora tiene valores de poco más de media hora. En términos relativos se puede observar que es solamente un 15-19% del tiempo medio, lo que indica que se están obteniendo predicciones bastante certeras.

A continuación se muestra en una tabla la ganancia en términos relativos:

Segmento	0-3º cuartil		Todos los datos		Ganancia precisión	Ganancia desviación
	Precisión	Desviación típica	Precisión	Desviación típica		
A	81,52%	15,36%	81,33%	25,20%	0,19%	9,84%
B	83,51%	13,59%	83,52%	23,45%	-0,01%	9,86%
C	82,78%	14,79%	82,49%	26,46%	0,29%	11,67%
D	83,53%	14,20%	84,55%	24,57%	-1,02%	10,37%
E	81,62%	15,59%	81,71%	29,22%	-0,09%	13,63%
F	77,56%	18,82%	77,82%	33,19%	-0,26%	14,37%
G	83,56%	14,95%	83,47%	24,85%	0,09%	9,90%
H	81,15%	16,89%	80,80%	28,38%	0,35%	11,49%
S	80,48%	16,66%	80,03%	31,57%	0,45%	14,91%
TA	81,10%	16,65%	82,29%	25,21%	-1,19%	8,56%
TB	80,64%	16,58%	80,61%	29,14%	0,03%	12,56%
TC	81,51%	15,69%	80,89%	29,89%	0,62%	14,20%
Promedio	81,58%	15,81%	81,63%	27,59%	-0,05%	11,78%

Tabla 37: Ganancia en horas de pintura por segmentar

No hay diferencias significativas en términos de precisión. Al igual que en los casos anteriores, la ganancia más importante se produce en la desviación típica, que ahora ha mejorado de media un 11,78%.

Por último, se muestra el ranking entre segmentos para la predicción de horas de mano de obra en pintura en el subconjunto 0-3º cuartil.

Orden absoluto	Error absoluto	Orden relativo	Precisión
A	0,58	G	83,56%
B	0,58	D	83,53%
G	0,61	B	83,51%
C	0,62	C	82,78%
D	0,64	E	81,62%
TA	0,65	A	81,52%
S	0,68	TC	81,51%
E	0,71	H	81,15%
H	0,72	TA	81,10%
TC	0,74	TB	80,64%
TB	0,74	S	80,48%
F	0,81	F	77,56%

Tabla 38: Ranking de segmentos en las horas de pintura para el subconjunto 0-3º cuartil

Los segmentos A, B, G y C son donde se obtienen predicciones con el menor error en horas, en contraste con el segmento F, que es donde más error hay. En términos relativos, se puede observar que en todos los segmentos se consiguen precisiones superiores al 80%, salvo en el segmento F con una precisión del 77%.

4.2.4 Horas de chapa

En este apartado se comentará los resultados obtenidos en la predicción de las horas de mano de obra en chapa de un siniestro.

Error Horas Chapa	A	B	C	D	E	F	G
Pintura, Reparación, Sustitución	1,76	1,60	1,67	1,66	1,91	2,19	1,81
Pintura, Reparación	2,47	2,43	2,42	2,40	2,72	3,17	2,51
Pintura, Sustitución	1,88	1,84	1,91	1,88	2,13	2,51	2,02
Reparación, Sustitución	1,79	1,82	1,79	1,87	1,99	2,29	1,92
Sustitución	2,33	2,24	2,19	2,24	2,42	2,89	2,38
Pintura	2,54	2,72	2,64	2,51	2,94	3,30	2,66
Reparación	3,09	3,28	2,83	2,93	3,23	3,75	3,00

Error Horas Chapa	H	S	TA	TB	TC	PROMEDIO
Pintura, Reparación, Sustitución	2,01	2,05	1,74	1,70	1,81	1,83
Pintura, Reparación	2,57	2,94	2,43	2,35	2,65	2,59
Pintura, Sustitución	2,16	2,28	1,97	2,00	2,04	2,05
Reparación, Sustitución	1,99	2,13	1,87	1,82	2,04	1,94
Sustitución	2,50	2,79	2,44	2,37	2,43	2,44
Pintura	2,79	3,28	2,58	2,48	2,73	2,76
Reparación	2,95	3,68	2,75	2,56	3,02	3,09

Tabla 39: Error absoluto en las horas de chapa

Desviación típica	A	B	C	D	E	F	G
Pintura, Reparación, Sustitución	2,35	2,27	2,27	2,37	2,99	3,22	2,31
Pintura, Reparación	3,37	3,80	3,66	3,54	4,46	4,80	3,58
Pintura, Sustitución	2,52	2,39	2,37	2,51	3,03	3,47	2,47
Reparación, Sustitución	2,48	2,35	2,34	2,43	3,03	3,43	2,47
Sustitución	2,69	2,59	2,59	2,71	3,27	3,52	2,73
Pintura	3,52	3,90	3,76	3,79	4,57	4,86	3,79
Reparación	4,11	4,35	4,32	4,22	5,19	5,36	4,35

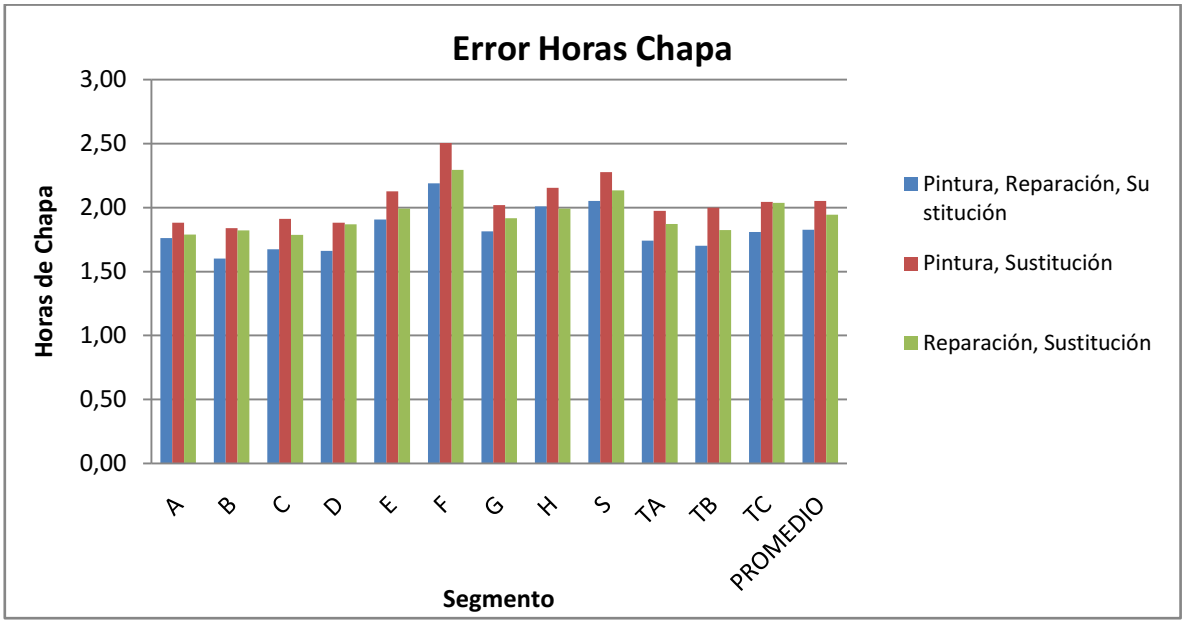
Desviación típica	H	S	TA	TB	TC	PROMEDIO
Pintura, Reparación, Sustitución	2,64	3,26	2,44	2,55	2,67	2,61
Pintura, Reparación	3,95	5,05	3,69	3,75	3,91	3,96
Pintura, Sustitución	2,82	3,30	2,58	2,61	2,80	2,74
Reparación, Sustitución	2,71	3,40	2,50	2,55	2,68	2,70
Sustitución	3,00	3,47	2,70	2,92	3,01	2,93
Pintura	4,07	4,98	3,84	3,87	4,11	4,09
Reparación	4,16	5,87	4,35	4,24	4,35	4,57

Tabla 40: Desviación típica de los errores en las horas de chapa

En estas dos tablas, se muestran los errores medios que se han obtenido en la predicción de las horas de mano de obra en chapa junto a sus desviaciones típicas.

Las redes neuronales que reciben solamente una variable como datos de entrada, junto al modelo [Pintura, Reparación] son los que mayor error medio tienen en sus predicciones. El peor modelo es [Reparación], con errores entre 2,5 y 3,75 horas, y con un error medio entre segmentos de 3,10 horas. Seguido de éste está [Pintura], con errores entre 2,5 y 3,3 horas, y con un error medio de 2,76 horas. Luego va el modelo conjunto de [Pintura, Reparación], con un error medio de 2,6 horas y finalmente está [Sustitución], que es el modelo que mejor predice individualmente, con un error medio de 2,44 horas. Se puede observar también que [Sustitución] es más preciso que el [Pintura, Reparación]; y que los modelos que mejor predicen, usan siempre como una de sus variables de entrada la sustitución de piezas.

A continuación, se muestran gráficamente los errores de las horas de pintura para los modelos que menor error tienen: [Pintura, Reparación Sustitución], [Pintura, Sustitución], y [Reparación, Sustitución].



Gráfica 8: Mejores modelos en horas de chapa

Tanto la gráfica 8 como la tabla 39 muestran que los modelos en el segmento F son donde se obtienen los mayores errores, seguido por los segmentos S, H y E. Por otra parte, los segmentos B, C y D son donde se obtienen los errores más pequeños.

En todos los segmentos salvo H, [Pintura, Reparación, Sustitución] es el modelo que más precisión tiene, con errores medios entre 1,6 y 2,2 horas y con una media ente todos los segmentos de 1,83 horas.

En el segmento H, el mejor modelo es [Reparación, Sustitución], con un error medio de 1,99 horas, que es prácticamente el mismo a las 2,01 horas obtenido en [Pintura, Reparación, Sustitución].

Si ordenamos de menor a mayor los segmentos según su error absoluto de predicción, se obtiene la siguiente lista:

Segmento	Error
B	1,60
D	1,66
C	1,67
TB	1,70
TA	1,74
A	1,76
TC	1,808
G	1,813
E	1,91
H	1,99
S	2,05
F	2,19

Tabla 41: Ranking de segmentos por error absoluto en horas de chapa

En cuanto a las desviaciones típicas mostradas en la tabla 40, éstas son generalmente mayores que los errores medios obtenidos. Por ejemplo, en el modelo [Pintura, Reparación, Sustitución], hay un error medio de 1,83 horas, y tiene una desviación típica promedia 2,61 horas.

Por otra parte, el ranking de los diferentes modelos de redes de neuronas según el error medio obtenido en todos los segmentos, queda de la siguiente manera.

Modelo	Error
Pintura, Reparación, Sustitución	1,83
Reparación, Sustitución	1,94
Pintura, Sustitución	2,05
Sustitución	2,44
Pintura, Reparación	2,59
Pintura	2,76
Reparación	3,09

Tabla 42: Ranking de modelos en horas de chapa por error absoluto

Se puede contemplar que el mejor modelo para predecir las horas de chapa es el que usa todas las variables: [Pintura, Reparación, Sustitución].

Análisis en términos relativos

Como en los casos anteriores, es interesante ver el error relativo medio de las predicciones, para tener un punto de comparación en el que basarse.

Precisión	A	B	C	D	E	F	G
Pintura, Reparación, Sustitución	63,80%	69,17%	66,80%	68,10%	66,14%	64,56%	67,22%
Pintura, Reparación	49,29%	53,23%	51,94%	53,94%	51,67%	48,76%	54,52%
Pintura, Sustitución	61,34%	64,62%	62,10%	63,85%	62,25%	59,47%	63,46%
Reparación, Sustitución	63,22%	64,95%	64,57%	64,11%	64,64%	62,87%	65,36%
Sustitución	52,20%	56,95%	56,48%	56,95%	56,99%	53,16%	57,04%
Pintura	47,70%	47,71%	47,69%	51,74%	47,72%	46,61%	51,81%
Reparación	36,44%	36,87%	43,87%	43,66%	42,71%	39,33%	45,84%

Precisión	H	S	TA	TB	TC	PROMEDIO
Pintura, Reparación, Sustitución	61,98%	65,54%	64,46%	65,26%	66,75%	65,81%
Pintura, Reparación	51,34%	50,56%	50,38%	52,01%	51,37%	51,58%
Pintura, Sustitución	59,23%	61,77%	59,71%	59,22%	62,41%	61,62%
Reparación, Sustitución	62,32%	64,15%	61,82%	62,78%	62,58%	63,61%
Sustitución	52,62%	53,14%	50,18%	51,72%	55,24%	54,39%
Pintura	47,26%	44,85%	47,43%	49,35%	49,84%	48,31%
Reparación	44,18%	38,15%	43,86%	47,74%	44,42%	42,26%

Tabla 43: Error relativo en horas de chapa

Desviación relativa	A	B	C	D	E	F	G
Pintura, Reparación, Sustitución	48,24%	43,78%	45,05%	45,50%	53,06%	52,11%	41,73%
Pintura, Reparación	69,34%	73,12%	72,50%	68,06%	79,18%	77,63%	64,78%
Pintura, Sustitución	51,84%	46,09%	47,07%	48,29%	53,82%	56,11%	44,61%
Reparación, Sustitución	50,91%	45,19%	46,39%	46,77%	53,77%	55,46%	44,66%
Sustitución	55,26%	49,88%	51,32%	52,07%	58,15%	56,96%	49,37%
Pintura	72,35%	75,13%	74,57%	72,76%	81,13%	78,67%	68,59%
Reparación	84,50%	83,73%	85,76%	81,01%	92,20%	86,68%	78,58%

Desviación relativa	H	S	TA	TB	TC	PROMEDIO
Pintura, Reparación, Sustitución	49,94%	54,68%	49,81%	52,04%	49,11%	48,75%
Pintura, Reparación	74,71%	84,82%	75,20%	76,56%	71,95%	73,99%
Pintura, Sustitución	53,38%	55,41%	52,67%	53,24%	51,53%	51,17%
Reparación, Sustitución	51,19%	57,12%	50,95%	51,97%	49,25%	50,30%
Sustitución	56,74%	58,22%	55,20%	59,51%	55,31%	54,83%
Pintura	76,93%	83,68%	78,39%	78,99%	75,58%	76,40%
Reparación	78,65%	98,50%	88,86%	86,46%	79,99%	85,41%

Tabla 44: Desviación típica relativa de los errores en las horas de chapa

En la tabla 42 se puede observar que en el mejor de los casos, las redes neuronales consiguen una precisión de entre el 60-70%, dependiendo del segmento, con una precisión media entre segmentos del 66%. Por otra parte, en la tabla 43 se ve que la desviación típica también es bastante alta siendo un 50% de la media de horas en los mejores modelos.

El modelo [Sustitución] es el que mejor resultados tiene individualmente consiguiendo una precisión media de 54,39%, pero al añadir las demás variables se obtienen una ganancia del 11%.

En la siguiente tabla se muestra el ranking de los segmentos por la precisión en sus predicciones. Se puede ver que relativamente el segmento donde mejor se predice es B D y G y en donde peor se predice A y H.

Segmento	Error
B	69,17%
D	68,10%
G	67,22%
C	66,80%
TC	66,75%
E	66,14%
S	65,54%
TB	62,26%
F	64,56%
TA	64,46%
A	63,80%
H	62,32%

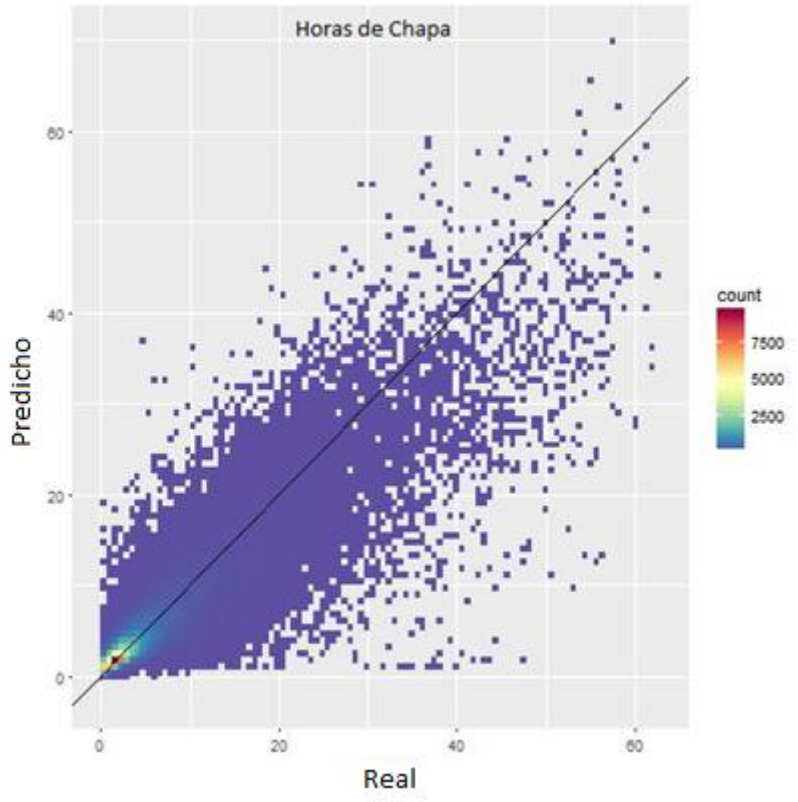
Tabla 45: Ranking de segmentos por precisión

Por otra parte, así queda el ranking de los modelos por error relativo:

Modelo	Error
Pintura, Reparación, Sustitución	65,81%
Reparación, Sustitución	63,61%
Pintura, Sustitución	61,62%
Sustitución	53,39%
Pintura, Reparación	51,58%
Pintura	48,31%
Reparación	42,26%

Tabla 46: Ranking de modelos de las horas de chapa por error relativo

A continuación se mostrará una gráfica de las predicciones en el segmento B. Se trata del mismo tipo de gráfico que los mostrados en los apartados anteriores, en donde se muestran los puntos (valor predicho, valor real), por lo que si el valor predicho y el valor real son el mismo, se situarán en la recta $y = x$, y cuanto más alejados de la recta estén, peor ha sido la predicción.



Gráfica 9: Predicción en horas de chapa

Nos encontramos con la misma situación que en los casos anteriores: hay un conjunto de puntos superpuestos en la gráfica, en donde se acumula la mayoría de los datos. Éstos precisamente coinciden al estar por debajo del 3º cuartil de las horas de chapa de los

siniestros. Por lo tanto, resulta interesante obtener modelos usando solamente en ese subconjunto de datos.

Resultados en el subconjunto mínimo-3ºcuartil

Segmento	Tiempo medio	Error medio	Desviación típica	Precisión	Desviación relativa
A	2,64	0,77	0,67	71,03%	25,30%
B	2,81	0,80	0,68	71,37%	24,26%
C	2,79	0,77	0,66	72,40%	23,57%
D	2,94	0,80	0,67	72,78%	22,96%
E	3,13	0,85	0,73	72,82%	23,15%
F	3,44	1,00	0,83	70,87%	24,08%
G	**3,13**	0,88	0,75	71,78%	24,04%
H	3,03	0,92	0,76	69,78%	25,05%
S	3,21	0,92	0,76	71,52%	23,73%
TA	2,80	0,86	0,68	69,12%	24,19%
TB	2,80	0,80	0,67	71,50%	23,99%
TC	3,08	0,87	0,72	71,59%	23,53%
Promedio	2,98	0,85	0,72	71,38%	23,99%

Tabla 47: Resultados del subconjunto 0-3º cuartil en horas chapa

Se puede observar que el error medio obtenido en cada segmento ha disminuido en casi dos horas aproximadamente: ahora se encuentra entre los 0,7-1 hora con una media de 0,85 horas, mientras que antes se encontraba entre las 1,60-2,2 horas y una media de 1,83 horas. La desviación típica también ha bajado bastante, antes estaba entre las 2,3-3,5 horas y una media de 2,61 horas, mientras que ahora tiene valores menores de una hora con una media de 0,72 horas.

Segmento	0-3º cuartil		Todos los datos		Ganancia precisión	Ganancia desviación
	Precisión	Desviación típica	Precisión	Desviación típica		
A	71,03%	25,30%	63,80%	48,24%	7,23%	22,94%
B	71,37%	24,26%	69,17%	43,78%	2,20%	19,52%
C	72,40%	23,57%	66,80%	45,05%	5,60%	21,48%
D	72,78%	22,96%	68,10%	45,50%	4,68%	22,54%
E	72,82%	23,15%	66,14%	53,06%	6,68%	29,91%
F	70,87%	24,08%	64,46%	52,11%	6,41%	28,03%
G	71,78%	24,04%	67,22%	41,73%	4,56%	17,69%
H	69,78%	25,05%	62,32%	51,19%	7,46%	26,14%
S	71,52%	23,73%	65,54%	54,68%	5,98%	30,95%
TA	69,12%	24,19%	64,46%	49,81%	4,66%	25,62%
TB	71,50%	23,99%	65,26%	52,04%	6,24%	28,05%
TC	71,59%	23,53%	66,75%	49,11%	4,84%	25,58%
Promedio	71,38%	23,99%	65,84%	48,86%	5,55%	24,87%

Tabla 48: Ganancia en horas de chapa por segmentar

En esta tabla se puede observar que términos relativos, se ha conseguido subir de media la precisión al 71%, mientras que antes se situaba en un 65%. Los segmentos que más ganancia han obtenido son A y H con un 7%, mientras que el que menos ha ganado es el segmento B, dado que ya tenía una precisión bastante alta. Se ha obtenido una ganancia media del 5,55%, aunque lo que más destaca es la ganancia en la desviación típica, que ha pasado de valores en torno al 50% a un 24% de media.

Finalmente, se muestra el ranking de segmentos ´para la predicción en el subconjunto de datos del 0 al 3º cuartil:

Orden absoluto	Error absoluto	Orden relativo	Precisión
A	0,77	E	72,82%
C	0,77	D	72,78%
B	0,80	C	72,40%
TB	0,80	G	71,78%
D	0,80	TC	71,59%
E	0,85	S	71,52%
G	0,88	TB	71,50%
TA	0,86	B	71,37%
TC	0,87	A	71,03%
H	0,92	F	70,87%
S	0,92	H	69,78%
F	1,00	TA	69,12%

Tabla 49: Ranking de segmentos en las horas de chapa para el subconjunto 0-3º cuartil

Los segmentos A, B, C son donde se obtienen predicciones con el menor error en horas, en contraste con el segmento F, que es donde más error hay. En términos relativos, se puede observar que en todos los segmentos se consiguen precisiones superiores al 70%, salvo en los segmentos H y TA, que prácticamente llegan al 70%.

4.1.5 Conclusión de los resultados obtenidos con redes neuronales

A lo largo del estudio con redes neuronales, se han probado 7 modelos para comprobar su capacidad de predicción. En la siguiente tabla se resumen los mejores modelos para cada variable a predecir:

Variable dependiente	Mejor modelo
Coste total	Pintura, Sustitución
Coste de piezas	Pintura, Sustitución
Horas de pintura	Pintura, Reparación, Sustitución
Horas de chapa	Pintura, Reparación, Sustitución

Tabla 50: Mejores modelos para cada variable dependiente

En la siguiente tabla se muestra un resumen de los resultados obtenidos:

Datos usados	Descriptivos	Coste total	Coste de piezas	Horas de pintura	Horas de chapa
Todos los datos	Error medio	299,55 €	206,34 €	1,1 horas	1,83 horas
	Desviación típica	458,18 €	384,78 €	1,62 horas	2,61 horas
	Precisión	73,02%	65,41%	81,51%	65,81%
0-3º cuartil	Error en	131,43 €	74,82 €	0,67 horas	0,85 horas
	Desviación típica	116,87 €	85,38 €	0,58 horas	0,72 horas
	Precisión	74,18%	68,00%	81,58%	71,38%

Tabla 51: Resumen de los resultados obtenidos

Las variables que mejor se predicen son las horas de pintura, con un acierto medio del 81% y un error medio de 1,1 horas por predicción al utilizar todos los datos, y prácticamente la misma precisión, y un error medio de 0,67 horas utilizando los datos de 0 al 3º cuartil. El coste total es la segunda variable que mejor se predice, con una precisión del 73% y del 74%, y unos errores medios de 300€ y 117€.

En tercer lugar está la predicción de las horas de chapa, en donde se han obtenido errores medios de 1,83 horas en todos los datos, y 0,85 horas en el subconjunto 0-3º cuartil, que traducido en tasa de acierto es un 65,81% y 71,38% respectivamente.

El peor resultado obtenido es la predicción del coste de las piezas, con un acierto medio del 65% en todos los datos, y 68% en el subconjunto 0-3º cuartil. De todas formas, la predicción de esta variable es la que menos importancia tiene, ya que generalmente a las aseguradoras les interesa más predecir el coste total, el cual en este caso tiene mejores resultados.

En todos los casos, las desviaciones típicas son mayores que los errores medios obtenidos, si se utilizan todos los datos. Esta alta desviación ha surgido debido a que las redes neuronales no han podido aprender a predecir las variables cuando éstos tienen valores extremadamente altos. A valores más altos, peor predicen, lo que ha originado esta gran dispersión en los errores.

Pero se ha podido observar que al coger un subconjunto de datos más pequeños, se ha disminuido notablemente la desviación típica de los errores, gracias a que las redes ya no tenían que lidiar con rangos tan altos de valores. Esto facilita la interpretación de los resultados, y agudiza la precisión de las predicciones.

Finalmente, en cuanto al ranking de predicciones por segmentos, éstos han variado según la variable a predecir, pero generalmente los segmentos A, B y C han sido donde se han obtenido las mejores predicciones, y F y H en donde se han obtenido las peores.

4.3 Modelos predictivos con LM

En este apartado se mostrarán los resultados obtenidos mediante los modelos lineales.

El estudio que se va a realizar es aplicable a todos los segmentos, por lo que en los siguientes apartados se analizará en detalle el segmento A, y se mostrarán los resultados de los demás segmentos en el Anexo B: Resultados del modelo lineal.

Se estimarán cuatro modelos de regresión, una para cada variable que se quiere predecir: coste total del siniestro, coste de las piezas, horas de mano de obra en pintura y horas de mano de obra en chapa, y también se explicará todo el proceso por el que se ha tenido que pasar para llegar a los modelos finales.

Finalmente, al igual que con las redes de neuronas, se probarán los modelos finales en el subconjunto de datos de 0 al 3º cuartil con el objetivo de mejorar lo máximo posible los resultados.

4.3.1 Consideraciones previas al análisis

Para determinar si es posible utilizar los LM, primero se ha tenido que estudiar que la distribución de probabilidad de la variable respuesta Y se adecúa a una distribución normal. Para estudiar las distribuciones de probabilidad se ha usado el test de Kolmogorov-Smirnov (KS-test), explicado en el apartado 2.7.4.

Cabe mencionar que al realizar un análisis preeliminar de los datos, solamente se consiguió ajustar los mismos a una distribución gamma. La razón de ello es que no se conocían los parámetros de la distribución teórica a la que se debía comparar la distribución empírica, y a que la distribución gamma ofrece gran versatilidad para ajustarse a los datos (Piboongungon, Aalo, Iskander y Efthymoglou, 2005). Por ello, al principio se optó por predecir estas variables mediante regresiones del modelo global GLM. Pero al realizar un estudio más detallado sobre las diferentes librerías de R, se descubrió una manera de buscar computacionalmente la distribución teórica a la que comparar los datos (Delignette-Muller, et al., 2016), y gracias a ello se consiguió demostrar que los datos siguen una distribución normal.

Basándonos en el criterio de simplicidad, se ha optado por utilizar el LM en vez del GLM, ya que el primero es menos complejo y a su vez más fácil de interpretar.

Estudio de distribuciones

A continuación se muestran los contrastes de bondad para la distribución normal en las variables a predecir: coste total del siniestro, coste de las piezas, horas de pintura y horas de chapa.

Segmento	Coste Total	Coste de las piezas	Horas Pintura	Horas de Chapa
A	20,45%	29,04%	14,11%	18,43%
B	20,81%	30,61%	13,98%	18,49%
C	21,98%	31,43%	15,59%	18,18%
D	21,34%	30,63%	16,82%	17,85%
E	23,91%	32,15%	17,26%	19,96%
F	23,63%	31,77%	18,57%	18,92%
G	19,62%	30,06%	15,66%	16,66%
H	20,24%	30,57%	16,92%	17,54%
S	24,91%	33,03%	15,99%	21,37%
TA	22,40%	30,95%	16,27%	18,95%
TB	23,33%	31,38%	18,07%	19,21%
TC	23,34%	31,32%	18,01%	18,70%

Tabla 52: KS-Test en los datos

En la tabla 52 se muestran los p-valor obtenidos de los contrastes de bondad de ajuste mediante el KS-Test. Usando la siguiente hipótesis:

$$H0: Y \text{ sigue una distribución normal}$$

$$H1: Y \text{ no sigue una distribución normal}$$

Se puede observar que en todos los casos, el p-valor que se obtiene está comprendido entre el 14-32%. Estableciendo un valor crítico usual del 5%, como el p-valor>5%, se rechaza la hipótesis alternativa frente la hipótesis nula. Esto que quiere decir que en todos los segmentos, la distribución de los datos y la distribución normal teórica correspondiente no difieren significativamente, por lo que se puede proceder a utilizar un LM para las predicciones.

Reducción de variables en el modelo

Para obtener una correcta interpretación en las regresiones, se han probado varios modelos antes de llegar al definitivo. Al principio se ha probado el modelo más simple, que es aquella en la que no se realiza ninguna transformación en los datos:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

Ecuación 37: Regresión con todas las variables

en donde Y es coste total, coste de piezas, horas de pintura y horas de chapa y $x_1, x_2, \ldots, x_k$ indica el número de piezas en cada zona del vehículo en donde se han realizado una de las 3 operaciones de reparación (sustitución, pintura, reparación). Cabe mencionar que en los modelos se han omitido aquellas variables que no eran significativas al 5%, es decir, aquellas que no aportaban información para predecir las variables dependientes. La salida de estos primeros modelos se puede encontrar en el Anexo B.

En todas las ocasiones, estos modelos presentaban un problema muy grave: muchos de los βs estimados son negativos, por ejemplo en la predicción del coste total, la pintura en la zona 3 tiene $\beta = -348{,}31$. Si interpretásemos este resultado, se diría que pintar una pieza más en la zona 3 del vehículo, disminuiría el coste total en 348,31€, lo cual no tiene sentido. Aquí se ha detectado un error de especificación en el modelo, lo que quiere decir que hay aspectos que no se están teniendo en cuenta en estas variables para predecir la variable dependiente. Este problema surge debido a la categorización extrema del vehículo en 27 zonas. Para solucionar este problema, se ha utilizado un segundo modelo en el que se agrupaban las zonas en zonas más grandes: las zonas 1-9 se han agrupado en una nueva zona 1, las zonas 10-18 en la zona 2, y las zonas 19-27 en la zona 3. Esto hace un total de 3 zonas, que combinado con los 3 métodos de reparación, se obtiene un modelo con 3x3=9 regresores:

$$Y = \beta_0 + \beta_1 s_1 + \beta_2 s_2 + \beta_3 s_3 + \beta_4 p_1 + \beta_5 p_2 + \beta_6 p_3 + \beta_7 r_1 + \beta_8 r_2 + \beta_9 r_3$$

Ecuación 39: Regresión con 9 variables

en donde s1, s2 y s3 son sustitución en las zonas 1, 2 y 3, p1, p2 y p3 es pintura en las mismas zonas, y r1, r2 y r3 reparación.

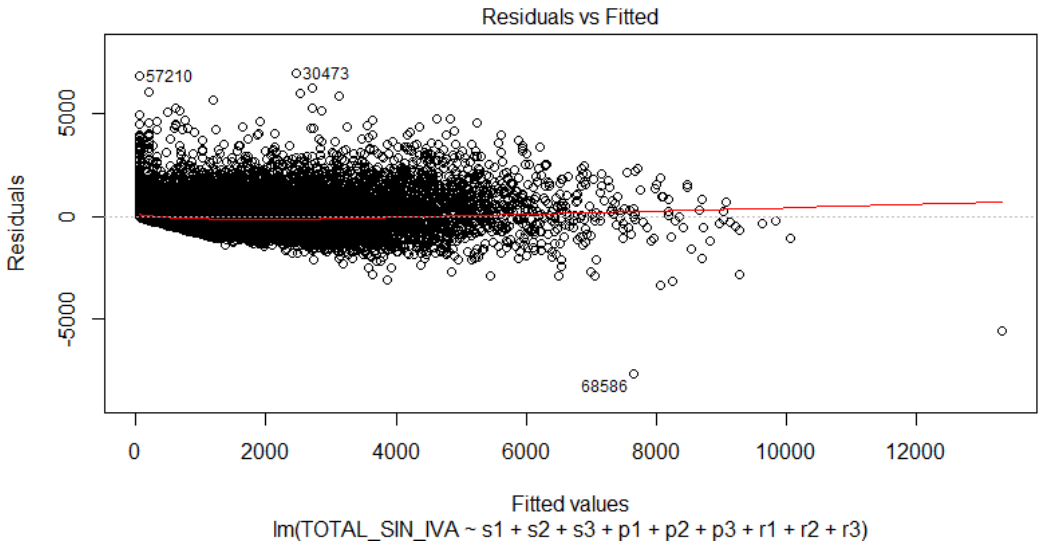
4.3.2 Coste total

En la siguiente tabla, se muestran los resultados obtenidos con la regresión de 9 variables para predecir el coste total:

Coeficientes	Estimación	Error estándar	Estadístico t	Pr(>|t|)	VIF
Intercepto	58,18	1,52	38,22	<2e-16	(–)
s1	139,81	0,36	385,59	<2e-16	1,89
s2	213,13	0,89	238,75	<2e-16	1,17
s3	129,25	0,62	207,47	<2e-16	1,77
p1	48,16	1,20	40,04	<2e-16	3,07
p2	22,07	0,99	22,20	<2e-16	2,66
p3	64,99	1,33	48,53	<2e-16	3,78
r1	12,40	1,67	7,40	1.33e-13	1,79
r2	24,69	1,72	14,27	<2e-16	2,07
r3	15,97	1,68	9,47	<2e-16	2,45
Error estándar de los residuos: 373,2 con 129348 grados de libertad					
R-cuadrado	0,8063	**R-cuadrado ajustado**	0,8063		
Estadístico F	5,98e+04	**p-valor:** 2,2e-16			
Durbin Watson	1,99	**p-valor**: 0,74			
Breusch Pagan	1515525	**p-valor:** 0			

Tabla 53: Regresión del coste total con 9 variables

Antes de poder interpretar el modelo, es necesario saber si se cumplen los supuestos del modelo lineal, en concreto, la linealidad, la multicolinealidad, homocedasticidad e independencia de errores. Para la linealidad, se analizará un gráfico de los residuos frente a los regresores.



Gráfica 10: Residuos frente a regresores en el coste total

En la gráfica 10 se puede observar que aunque haya una gran cantidad de puntos superpuestos entre sí, los puntos siguen la recta roja, lo que indica que tiene una tendencia es lineal. Con la linealidad en los parámetros se puede interpretar que un incremento unitario en X_j tiene el mismo efecto sobre el coste total con independencia del valor inicial de X_j (manteniendo los demás regresores constantes).

Como se ha explicado en el apartado 2.7.2, para la multicolinealidad, homocedasticidad e independencia de errores se ha usado el factor de inflación de la varianza, el test de Breusch Pagan, y el test de Dubin Watson respectivamente.

En la tabla 53 se puede observar que los factores de inflación de varianza tienen valores entre 1,1 y 3,7, lo que indica que hay poca multicolinealidad entre las variables. Las variables que más multicolinealidad presentan son los de pintura. Una alternativa para disminuir estos factores de inflación de la varianza, sería hacer un análisis de componentes principales en estas variables. Este análisis no se ha realizado debido a que estos valores no son altos, y son menores de 10.

Respecto al test de independencia de Dubin Watson, el p-valor es mayor del 5%. Siguiendo la hipótesis presentado en el apartado 2.8.2.3:

$$H0: no\ existe\ autocorrelación$$

$$H1: existe\ autocorrelación$$

Se rechaza la hipótesis alternativa frente a la nula, lo que indica que no existe autocorrelación entre los errores y por ello, son independientes.

El problema está con el segundo test, el test de Breusch Pagan para la homocedasticidad. Con el contraste de hipótesis presentado en el apartado 2.8.2.2:

$$H0: El\ modelo\ es\ homomocedástico$$

$$H1: El\ modelo\ es\ heterocedástico$$

Dado que el p-valor es 0, se rechaza la hipótesis nula de homocedasticidad, lo que quiere decir que la varianza no es constante y por ello es heterocedástica. Esto quiere tal y como se ha explicado en el apartado 2.7.2.2, los parámetros obtenidos siguen siendo insesgados y consistentes, por lo que se pueden usar sin problemas, pero los errores estándar de las estimaciones son sesgados e inconsistentes, y esto invalida la inferencia de los estimadores (Goldberger, 1991).

Como que el único problema son los errores estándar, la solución más común es usar los errores estándar de Huber-White que son robustos a la heterocedasticidad. A continuación se muestra la regresión con estimadores robustos en los errores estándar:

Coeficientes	Estimación	Error robusto	Estadístico t	Pr(>|t|)
Intercepto	58,18	1,89	31,26	<2e-16
s1	139,81	0,77	179,48	<2e-16
s2	213,13	2,63	81,39	<2e-16
s3	129,25	1,24	104,30	<2e-16
p1	48,16	1,69	28,31	<2e-16
p2	22,07	1,33	15,71	<2e-16
p3	64,99	1,76	35,45	<2e-16
r1	12,40	2,07	6,49	8.65e-11
r2	24,69	2,24	11,14	<2e-16
r3	15,97	2,15	6,78	1.17e-11
Estadístico F	1,29e+07	p-valor: <2,2e-16		
Error medio	203,05€	**Desviación típica**	271,05€	
Media del segmento	716,00€	**Precisión del modelo**	71,64%	

Tabla 54: Regresión en el coste total con errores estándar robustos

Se puede observar en la tabla 54 que los errores estándar robustos aumentan en comparación a los errores estándar de la Tabla 53, y en ocasiones hasta más del doble como es el caso de s2. Esto se puede considerar el precio a pagar por usar unos errores robustos a la heterocedasticidad. De todas formas, éstos siguen siendo bastante pequeños en comparación a sus estimadores.

Al cambiar los errores robustos, el cálculo del estadístico t varía un poco, pero todas las variables siguen siendo significativas al 1%, lo que indica que los 3 tipos de reparación en estas 9 zonas influyen significativamente en el coste del siniestro y el modelo está bien especificado.

Respecto al estadístico F, éste sirve para analizar globalmente la regresión, comparando la discrepancia que hay entre un modelo de regresión que no tiene variables predictoras, es decir, que solamente usa el intercepto, con el modelo de regresión especificado. Para ello se utiliza el siguiente contraste de hipótesis:

$$H0: El\ modelo\ con\ solo\ el\ intercepto\ y\ el\ modelo\ especificado\ son\ iguales$$

$$H1: El\ ajuste\ del\ modelo\ del\ intercepto\ es\ significativamente\ peor\ que\ el\ modelo\ especificado$$

Como el p-valor del estadístico F es prácticamente 0, se rechaza la hipótesis nula a un nivel de significación del 1%. Con esto se concluye que el modelo especificado está mejor ajustado que el modelo con solo el intercepto. los βs estimados son en conjunto necesarios para predecir el coste total.

La variable que más influye en el coste total del siniestro es la sustitución de piezas. En concreto, en la zona 2, si se sustituye una pieza adicional y se mantienen las demás variables constantes, el coste aumenta en promedio 213€. La sustitución en las zonas 1 y 3 aportan un coste entre 130-140€ por pieza adicional.

Las siguientes variables que más influyen son la pintura en la zona 1 y 3, que aumentan unos 50-65€ por cada pieza adicional que se pinte. Por último, las variables que menos influyen en el coste total son la reparación y la pintura en la zona 2, con aumentos del coste entre 12-25€.

Estos resultados concuerdan con los obtenidos en las redes de neuronas en el apartado 4.2.1, en donde se vio que los modelos que usan la variable sustitución, son las que mejor predicen; que la pintura y la reparación combinadas aportan menos información que sustitución, y que la reparación de piezas es la variable que menos influye en el coste total.

Tal y como se mencionó en el apartado 2.7.2.2, en presencia de heterocedasticidad, no tiene sentido considerar el error estándar de la regresión o el R^2, puesto que sólo tienen sentido en un contexto homocedástico (Goldberger, 1991). Como alternativa a la medición de bondad del modelo, se utilizará la precisión, definida en la ecuación 35, que se obtiene midiendo de los errores de las predicciones de un conjunto de datos para testear el modelo. Se puede observar que en este caso se obtiene una precisión del 71,64%, es decir, de media este modelo se está equivocando un 29% en cada predicción, lo que equivale a 203€ en términos monetarios. Estos errores tienen una desviación típica de 271,05€, el cual es bastante alto, ya que supera incluso al error medio obtenido. Esto indica que hay una gran dispersión en los errores obtenidos, lo que dificulta la fiabilidad de la predicción.

4.3.3 Coste de las piezas

Cabe mencionar, que en el caso del coste de las piezas, el modelo obtenido de 9 variables seguía sin tener una correcta especificación: se obtenían estimaciones negativas, el cual no tiene sentido ya que el aumento de trabajo no disminuye su coste. Por ello, se ha tenido que reducir todavía más el modelo, a uno en el que se usan 3 variables, correspondientes a la sustitución, reparación de piezas y pintura de piezas, sin tener en cuenta las zonas en donde realizan:

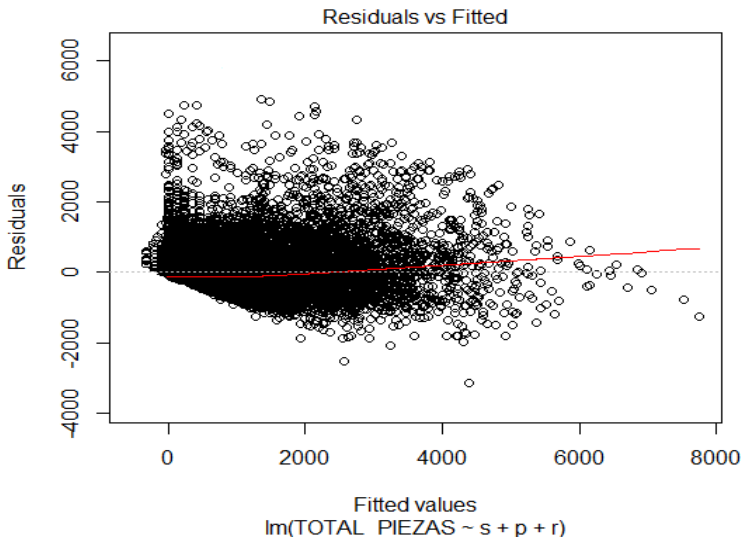
En la siguiente tabla, se muestran los resultados obtenidos con la regresión de 3 variables para predecir el coste de piezas:

Coeficientes	Estimación	Error estándar	Estadístico t	Pr(>|t|)	VIF
Intercepto	12,06	1,06	11,33	4,71e-05	(–)
Sustitución	112,66	0,17	629,93	<2e-16	1,08
Pintura	7,35	0,59	8,46	<2e-16	1,91
Reparación	2,73	0,32	12,40	<2e-16	1,88
Error estándar de los residuos: 303,5 con 109882 grados de libertad					
R-cuadrado	0,7585	**R-cuadrado ajustado**	0,7585		
Estadístico F	1,15e+05	**p-valor:** 2,2e-16			
Durbin Watson	1,99	**p-valor**: 0,664			
Breusch Pagan	1515525	**p-valor:** 0			

Tabla 55: Regresión del coste de piezas con 3 variables

Al igual que en el caso anterior, se estudiará el cumplimiento de supuestos de linealidad, multicolinealidad, homocedasticidad e independencia para ver si se puede interpretar este modelo correctamente.

A continuación se muestra el gráfico de los residuos frente a los regresores para estudiar la linealidad del modelo.



Gráfica 11: Residuos frente a regresores en el coste de piezas

Al igual que antes, hay una gran cantidad de puntos superpuestos entre sí en la que se puede observar que la tendencia de las mismas es aproximadamente lineal, por ello, no hay indicios de no linealidad en esta gráfica.

Respecto al test de multicolinealidad, en la tabla 55 muestra que el factor de inflación de varianza tiene valores entre 1 y 2, lo que indica escasa multicolinealidad. Este resultado es obvio ya que se ha reducido una gran cantidad de variables a un total de 3, por lo que la multicolinealidad ha disminuido casi por completo.

En cuanto al test de independencia de Dubin Watson, el p-valor es 0,66, que es mayor que el 5%, lo que indica independencia de los errores. El problema sigue estando en el test de Breusch Pagan, con un p-valor igual a 0, que indica que la varianza de los errores no es constante. Para lidiar con la heterocedasticidad, se ha usado también el modelo lineal con errores estándar robustos a la heterocedasticidad:

Coeficientes	Estimación	Error robusto	Estadístico t	Pr(>|t|)
Intercepto	12,06	1,21	9,91	<2e-16
Sustitución	112,66	0,50	225,31	<2e-16
Pintura	7,35	0,64	6,83	8,35e-12
Reparación	2,73	0,39	11,47	<2e-16
Estadístico F	1,29e+07	p-valor: <2,2e-16		
Error medio	145,74€	**Desviación típica**	223,81€	
Media del segmento	378,48€	**Precisión del modelo**	61,49%	

Tabla 56: Regresión en el coste de piezas con errores estándar robustos

Tal y como se muestra en esta tabla, todos los regresores obtenidos son significativos, lo que indica que el modelo está bien especificado y no hay ninguna variable que no influya en el coste de piezas. Los errores estándar robustos han aumentado frente a los mostrados en la Tabla 55, y entre ellos destaca un aumento de casi el triple en el error estándar robusto de la sustitución respecto al anterior, pasando de 0,17 a 0,5, aunque como en el coste total, éstos siguen siendo pequeños en comparación con los estimadores.

Respecto al estadístico F, el p-valor es prácticamente 0, por lo que se rechaza la hipótesis nula a un nivel de significación del 1% y se concluye que el modelo especificado está mejor ajustado que el modelo que usa solamente el intercepto.

La variable que más influye en el coste de las piezas es de lejos la sustitución, lo cual es obvio ya que el número total de piezas sustituidas es directamente proporcional al coste de las mismas. En concreto, si se sustituye una pieza adicional, manteniendo las demás variables constantes, el coste de las piezas aumenta en promedio 112,66€.

Por otra parte, la pintura y la reparación influyen mucho menos y pintar una pieza adicional aumenta el coste en 7,35€ de media, y reparar una pieza 2,73€, es decir, que entre ambos influyen menos del 10% de lo que aporta la sustitución de piezas.

Como que en presencia de heterocedasticidad, no tiene sentido considerar el error estándar de la regresión o el R^2, para medir la bondad del modelo se ha usado la precisión. Se puede observar que en este caso se obtiene una precisión del 61,49%, es decir, de media este modelo se está equivocando un 39% en cada predicción, lo que equivale a 145,74€ en términos monetarios. Estos errores tienen una desviación típica de 223,81€, lo que añade bastante incertidumbre en las predicciones obtenidas.

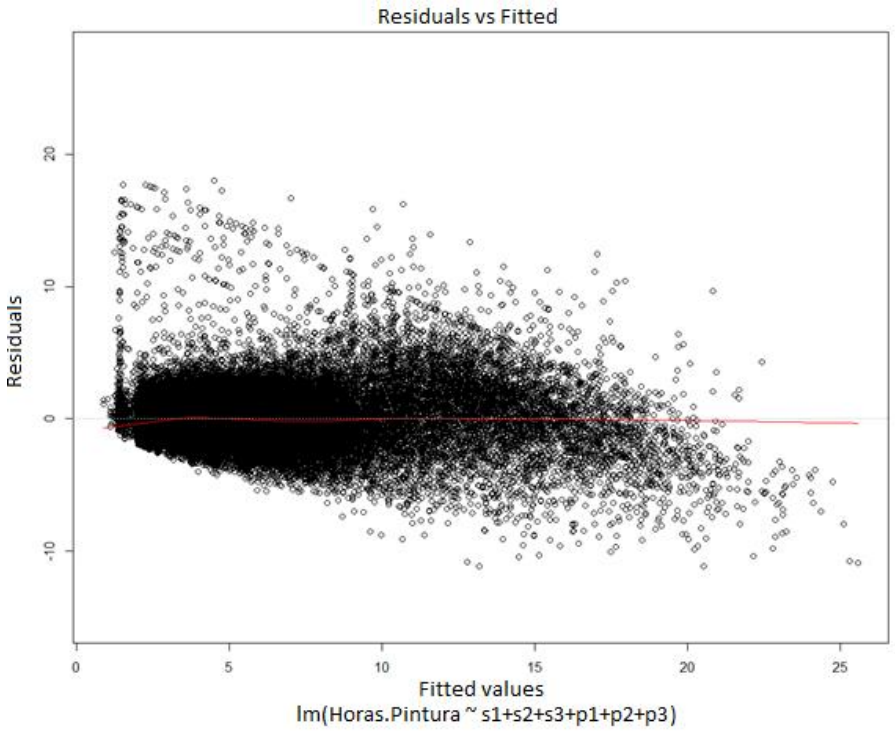
4.3.4 Horas de pintura

En la siguiente tabla, se muestran los resultados obtenidos con la regresión de 9 variables para predecir las horas de pintura:

Coeficientes	Estimación	Error estándar	Estadístico t	Pr(>|t|)	VIF
Intercepto	1,41	0,0079	178,69	<2e-16	(–)
s1	0,02	0,0016	9,83	<2e-16	1,90
s2	0,15	0,004	38,93	<2e-16	1,16
s3	0,04	0,0027	15,48	<2e-16	1,75
p1	0,81	0,0054	148,98	<2e-16	3,10
p2	0,66	0,0044	152,29	<2e-16	2,65
p3	0,57	0,006	94,92	<2e-16	3,77
r1	0,21	0,0073	29,08	1.33e-13	1,79
r2	0,43	0,0075	57,13	<2e-16	2,06
r3	0,43	0,0074	58,11	<2e-16	2,45
Error estándar de los residuos: 1.582 con 105079 grados de libertad					
R-cuadrado	0,7694	**R-cuadrado ajustado**	0,7694		
Estadístico F	3,89e+04	**p-valor:** 2,2e-16			
Durbin Watson	1,99	**p-valor**: 0,312			
Breusch Pagan	84954,01	**p-valor:** 0			

Tabla 57: Regresión de las horas de pintura con 9 variables

Se puede observar que reduciendo el modelo a 9 variables es suficiente para obtener una especificación correcta. A continuación se muestra la gráfica de residuos frente a los regresores para comprobar la linealidad del modelo:



Gráfica 12: Residuos frente a regresores en las horas de pintura

En la Gráfica 12 se puede observar que hay algunos puntos en la zona superior izquierda de la gráfica que podrían indicar no linealidad, pero son muy pocos comparados con los que siguen la línea roja. Por ello, se considera que este modelo también es lineal en los parámetros.

Respecto a los test de multicolinealidad, independencia y homocedasticidad se obtienen resultados similares a los anteriores: el factor de inflación de varianza tiene valores entre 1,16 y 3,77, lo que indica que hay algo de multicolinealidad entre las variables, pero poca, por lo que no hay motivo para hacer un análisis de componentes principales. El test de independencia de Dubin Watson da como resultado un p-valor de 0,31, que es mayor que 0,05, lo que indica independencia de los errores. El problema sigue estando en el test de Breusch Pagan de varianza constante, en la que se da como resultado la heterocedasticidad. Por ello se ha usado los errores estándar robustos a la heterocedasticidad:

Coeficientes	Estimación	Error robusto	Estadístico t	Pr(>|t|)
Intercepto	1,41	0,009	156,63	<2e-16
s1	0,02	0,002	7,38	<2e-16
s2	0,15	0,006	23,40	<2e-16
s3	0,04	0,0038	10,59	<2e-16
p1	0,81	0,0078	103,76	<2e-16
p2	0,66	0,0078	84,97	<2e-16
p3	0,57	0,0093	61,54	<2e-16
r1	0,21	0,0106	20,04	1.33e-13
r2	0,43	0,0133	30,99	<2e-16
r3	0,43	0,0108	39,27	<2e-16
Estadístico F	1,29e+07	p-valor: <2,2e-16		
Error medio	0,98	**Desviación típica**	1,11	
Media del segmento	4,58	**Precisión del modelo**	78,58%	

Tabla 58: Regresión en las horas de pintura con errores estándar robustos

En esta tabla se puede observar que tal y como sucede en los casos anteriores, los errores estándar robustos aumentan bastante en comparación a los errores estándar mostrados en la tabla 57.

Todas los regresores obtenidos son significativos, lo que indica que el modelo está bien especificado y no hay ninguna variable que no influya en la estimación de las horas de mano de obra en pintura.

Respecto al estadístico F, el p-valor es prácticamente 0, por lo que se rechaza la hipótesis nula a un nivel de significación del 1% y se concluye que el modelo especificado está mejor ajustado que el modelo que usa solamente el intercepto.

La variable que más influye en las horas de pintura es la pintura, lo cual tiene su explicación, ya que el número total de piezas pintadas es directamente proporcional a las horas necesarias para pintarlas. En concreto, si se pinta una pieza adicional, en media se requiere entre 0.6-0.8 horas adicionales de mano de obra.

Por otra parte, la sustitución y la reparación influyen mucho menos. Sustituir una pieza adicional apenas aumenta las horas en pintura, lo cual es lógico ya que normalmente no es necesario pintar una pieza nueva, salvo que se quiera cambiar el color de la chapa del automóvil. Reparar una pieza adicional, requiere un aumento de 0.2-0.45 horas adicionales de mano de obra. En conclusión, la pintura es la variable que más influye, seguido de la reparación, y por último está la sustitución.

Se puede observar que en este modelo, se obtiene una precisión del 78,58%, es decir, de media este modelo se está equivocando un 22% en cada predicción, lo que equivale a una hora de mano de obra. Estos errores tienen una desviación típica de 1,11 horas, lo que añade algo de incertidumbre en las predicciones obtenidas.

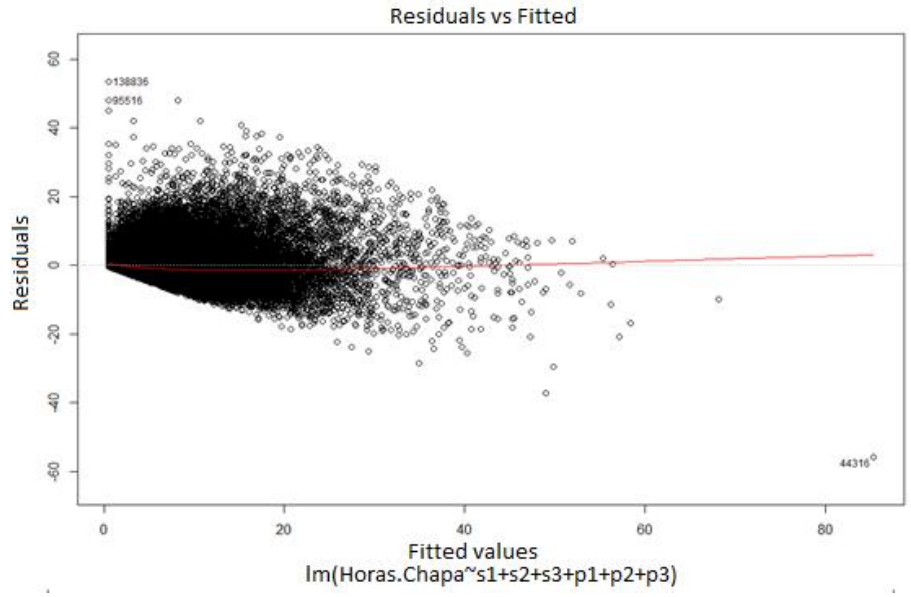
4.3.5 Horas de chapa

En la siguiente tabla, se muestran los resultados obtenidos con la regresión de 9 variables para predecir las horas de chapa:

Coeficientes	Estimación	Error estándar	Estadístico t	Pr(>|t|)	VIF
Intercepto	0,50	0,0146	34,491	<2e-16	(–)
s1	0,71	0,0034	209,809	<2e-16	1,89
s2	1,34	0,0083	160,584	<2e-16	1,17
s3	0,93	0,0058	160,535	<2e-16	1,75
p1	0,021	0,0113	5,893	0,0161	3,07
p2	0,019	0,0093	6,082	0,0274	2,66
p3	0,32	0,0126	25,403	<2e-16	3,76
r1	0,69	0,0157	43,849	<2e-16	1,79
r2	0,55	0,0162	34,237	<2e-16	2,06
r3	0,75	0,0158	47,754	<2e-16	2,44
Error estándar de los residuos: 1.582 con 105079 grados de libertad					
R-cuadrado	0,6306	**R-cuadrado ajustado**	0,6306		
Estadístico F	3,89e+04	**p-valor:** 2,2e-16			
Durbin Watson	2,03	**p-valor**: 0,518			
Breusch Pagan	136553	**p-valor:** 0			

Tabla 59: Regresión de las horas de chapa con 9 variables

El reducir el modelo a 9 variables es suficiente para obtener una especificación correcta. A continuación se muestra la gráfica de residuos y estimadores para comprobar la linealidad del modelo:



Gráfica 13: Residuos frente a regresores en las horas de chapa

Como en los casos anteriores, se puede observar en la gráfica 13 que la nube de puntos tiene una tendencia de seguir la recta roja, lo que indica linealidad en el modelo.

Respecto a los test de multicolinealidad, independencia y homocedasticidad se obtienen resultados similares a los anteriores: el factor de inflación de la varianza es se encuentra entre valores del 1,1 y el 3,8, lo que indica que hay un poco de multicolinealidad entre las variables, pero no mucho. En el test de independencia de Dubin Watson, el p-valor es 51,8%, que es mayor que el 5%, lo que indica independencia de los errores. El problema sigue estando en el test de Breusch Pagan, que indica que la varianza de los errores no es constante. Para lidiar con ello, se ha usado los errores estándar robustos.

Coeficientes	Estimación	Error robusto	Estadístico t	Pr(>|t|)
Intercepto	0,50	0,0172	29,121	<2e-16
s1	0,71	0,007	101,506	<2e-16
s2	1,34	0,026	51,681	<2e-16
s3	0,93	0,0123	76,323	<2e-16
p1	0,021	0,0156	1,373	0,027
p2	0,019	0,1334	1,449	0,034
p3	0,32	0,177	18,136	<2e-16
r1	0,69	0,02	34,588	<2e-16
r2	0,55	0,0244	22,754	<2e-16
r3	0,75	0,022	33,05	<2e-16
Estadístico F	6159	p-valor: <2,2e-16		
Error medio	2,02	**Desviación típica**	2,72	
Media del segmento	4,86	**Precisión del modelo**	58,42%	

Tabla 60: Regresión en las horas de chapa con errores estándar robustos

En esta tabla se puede observar que todos los regresores obtenidos son significativos al 1% salvo p1 y p2, que lo son al 5%. Esto indica que el modelo está bien especificado y no hay ninguna variable que no influya en estimación de las horas de mano de obra en chapa.

Respecto al estadístico F, el p-valor es prácticamente 0, por lo que se rechaza la hipótesis nula a un nivel de significación del 1% y se concluye que el modelo especificado está mejor ajustado que el modelo que usa solamente el intercepto.

Se puede observar que la variable que más influye en las horas de chapa es la sustitución de piezas. Si se sustituye una pieza adicional, en media se requiere entre 0.7-1.34 horas adicionales de mano de obra.

La segunda variable que más influye en esta regresión es la reparación de piezas, en la que sustituir una pieza adicional, requiere entre 0.55-0.75 horas adicionales en mano de obra de chapa.

Por último, la variable que menos influye es la pintura de piezas, en la que pintar una pieza adicional aporta entre 0.06-0.32 horas.

Al igual que en los casos anteriores, para medir la bondad del modelo se ha utilizado la precisión. Se puede observar que en este caso se obtiene una precisión del 58,42%, es decir, de media este modelo se está equivocando un 42% en cada predicción, lo que equivale a 2

horas de mano de obra. Este error se considera bastante alto, y empeora con la alta desviación típica de 2,72 horas.

4.3.6 Modelo lineal en el 0-3º cuartil

Al igual que sucede en los modelos de redes neuronales, con el objetivo de mejorar los resultados y sobretodo disminuir la desviación típica obtenida en los errores de predicción, se probará el modelo lineal en el subconjunto de datos correspondiente a los tres primeros cuartiles. Estos modelos tienen las mismas características que los presentados en los apartados anteriores: se cumplen todos los supuestos salvo el de la homocedasticidad, por lo que se han usado errores robustos. Los resultados están presentados en el Anexo B: Resultados del modelo lineal.

En la siguiente tabla se muestran los parámetros estimados, junto al error medio, su desviación típica y la precisión de las regresiones realizadas en las cuatro variables que se quieren predecir:

Coeficientes	Coste total	Coste de piezas	Horas de pintura	Horas de chapa
Intercepto	145,9	56,26	1,8	1,29
s1	66,11	42,48	0,006	0,28
s2	77,07		0,095	0,36
s3	56,33		0,013	0,17
p1	44,13	1,58	0,452	0,07
p2	41,73		0,516	0,18
p3	39,07		0,389	0,02
r1	28,46	1,04	0,298	0,51
r2	19,42		0,231	0,21
r3	42,94		0,46	0,71
Error medio	96,83 €	63,46 €	0,72	0,89
Desviación típica	81,34 €	52,24 €	0,52	0,73
Precisión	73,69%	59,32%	76,17%	66,41%

Tabla 61: Modelo lineal en el subconjunto 0-3º cuartil

Para el coste total, las variables más significativas son la sustitución de piezas, que aporta entre 56-77€ por pieza adicional. Ésta seguida por la pintura, que aporta entre 39-44€ por pieza pintada adicional. Y por último está la reparación, en donde influye entre 19-42€ en el coste. En este modelos se ha obtenido un error medio e 96,83€ y una precisión del 74%, frente a los 203€ y 72% de precisión obtenido usando todos los datos. La diferencia que había en el impacto de la sustitución frente a la pintura o reparación se ha reducido en este subconjunto de datos. Esto indica que al tener en cuenta siniestros con costes muy altos, se opta directamente por sustituir las piezas en vez de pintarlas o repararlas.

Respecto al coste de piezas, la variable que más influye sigue siendo la sustitución de piezas, que aporta 42€ por pieza adicional. Se ha obtenido en este modelo de regresión una

precisión del 60% junto a un error medio de 63€, y una desviación típica del error de 52€, frente al error medio de 146€ y desviación de 224€ obtenidos al usar todos los datos.

En la regresión de las horas de pintura, la variable más significativa sigue siendo pintura: cada pieza adicional pintada requiere entre 0,38-0,51 horas de mano de obra. La segunda más importante es la reparación, y por último la sustitución. Aquí se ha obtenido un error medio de 0,72 horas, junto a una desviación típica de 0,52 horas, y una precisión del 76,17%. Aunque el error y la desviación típica se haya reducido en términos absolutos, la precisión ha disminuido un poco, ya que al usar todos los datos se ha conseguido un 78%.

Por último, en la regresión de horas de chapa, se puede observar que ahora la variable que más influye es la reparación de piezas en vez de la sustitución. La primera aporta 0,2-0,7 horas de mano de obra por pieza adicional, y la sustitución 0,17-0,36 horas. La variable que menos influye por lo tanto es la pintura, que aporta de media entre 0,02-0,51 horas de mano de obra por pieza. En este modelo se ha obtenido un error medio de 0,89 horas y 0,73 horas de desviación típica, que se traduce a un 66,41% en precisión. Esto ha supuesto una gran mejoría, ya que usando todos los datos se obtuvo solamente un 58,42% de precisión.

4.3.7 Conclusiones del modelo lineal

En la tabla 62 se puede observar que se han obtenido en general buenos resultados en el coste total con una precisión del 72%, y en las horas de pintura con un 78%. Respecto a la predicción del coste de piezas y horas de chapa, se han obtenido una tasa de aciertos del 61% y 58% respectivamente, los cuales se consideran resultados bastante mediocres.

Coeficientes	Coste total	Coste de piezas	Horas de pintura	Horas de chapa
Intercepto	58,18	12,06	1,41	0,5
s1	139,81	112,66	0,02	0,71
s2	213,13		0,15	1,34
s3	129,25		0,04	0,93
p1	48,16	7,35	0,81	0,021
p2	22,07		0,66	0,019
p3	64,99		0,57	0,32
r1	12,4	2,73	0,21	0,69
r2	24,69		0,43	0,55
r3	15,97		0,43	0,75
Error medio	203,05	145,77 €	0,98	2,02
Desviación típica	271,05	224 €	1,10	2,73
Precisión	71,64%	61,49%	78,58%	58,42%

Tabla 62: Comparación de los modelos lineales usando todos los datos

Respecto a los βs obtenidos en los 4 modelos de regresión se puede observar que tanto para las estimaciones del coste total y el coste de las piezas, la variable que más influye en la predicción es la sustitución de piezas. seguida de la pintura, y quedando en última posición la reparación. En la estimación de las horas de chapa, la sustitución también es la variable que

más influye, pero ésta es seguida de la reparación, y la pintura apenas influye. Pero se ha visto en la tabla 61 que cuando se utiliza solamente los datos del 0 al 3º cuartil, la reparación es la que más influye.

Por último, en la predicción de las horas de pintura, la variable que más influye es la pintura de piezas, seguido de la reparación y por último la sustitución.

Se ha visto que se han tenido que pasar por múltiples modelos hasta llegar a las regresiones finales. Cabe destacar que el modelo lineal requiere supuestos bastante estrictos y los resultados obtenidos en este proyecto es otro ejemplo de la vida real en el que no éstos se cumplen completamente. En este caso, los modelos obtenidos no son homocedásticos, por lo que no ha sido posible obtener los estimadores lineales insesgados de mínima varianza.

También se ha podido observar que existen remedios para lidiar con el incumplimiento de los supuestos. La solución en este caso ha sido utilizar estimadores robustos de los errores estándar, pero utilizarlos han conllevado un coste, y es el de obtener un error en ocasiones varias veces mayor. Aunque en este caso no haya influido tanto, dado que los errores estándar de las estimaciones eran muy pequeñas en comparación a las βs, no se garantiza que esto suceda en otras situaciones.

4.4 Comparación entre redes de neuronas y modelo lineal

A continuación se muestra una tabla resumiendo la precisión media obtenida en ambos modelos para las diferentes regresiones. Los resultados obtenidos en cada segmento utilizando el modelo lineal se encuentran en el Anexo B.

Variable dependiente	Datos utilizados	LM	Redes de neuronas
Coste total	Todos los datos	71,92%	73,02%
	0-3º cuartil	70,55%	74,18%
Coste piezas	Todos los datos	56,50%	65,41%
	0-3º cuartil	55,82%	68,00%
Horas de pintura	Todos los datos	78,41%	81,51%
	0-3º cuartil	78,18%	81,58%
Horas de chapa	Todos los datos	61,16%	65,81%
	0-3º cuartil	68,19%	71,38%

Tabla 63: Comparación de modelo lineal y redes de neuronas

Se puede observar que en general se obtienen mejores resultados al utilizar redes neuronales. Para el coste total, horas de pintura y horas de chapa, suelen ser un 3% más precisos, por lo que no es mucho. La mayor diferencia se produce en la predicción del coste de piezas, en donde las precisiones de las redes neuronales son un 10-12% superiores a las del LM. Esto es indicio de que la red neuronal ha conseguido aprender algún patrón oculto de los datos que no ha detectado el modelo lineal, por lo que se entiende que éste tiene mayor capacidad de resolver problemas más complejos.

Existen varias diferencias entre los modelos lineales y redes neuronales

1. En los modelos estadísticos, se requiere realizar un estudio intensivo de los datos antes de poder utilizar los modelos: es necesario conocer qué tipo de distribución siguen los datos, y si se utiliza una distribución poblacional errónea, los resultados obtenidos son pésimos. Este análisis previo no es necesario en las redes neuronales.
2. Los modelos lineales requieren que se cumplan una serie de supuestos para la correcta especificación del modelo, y en ocasiones, estos supuestos pueden ser bastante estrictos. Si no se cumplen, los resultados obtenidos son pobres. En contraste, las redes neuronales tampoco requiere que se cumplan supuestos previos para el análisis de los datos.
3. Por otra parte, el uso de los LM es menos automatizado: tras crear los modelos, es necesario ir comprobando paso a paso que se cumplen todos los supuestos. Si éstos no se cumplen, se tiene que buscar una manera de solucionar el problema, osino, no es posible interpretar correctamente el modelo.
4. No son todo desventajas en los modelos lineales. Éstos ofrecen un enfoque distinto, y se centran bastante en inferir de manera correcta los parámetros, mientras que en las redes neuronales el objetivo más importante es el de la predicción. Se ha podido comprobar que con las redes neuronales es bastante más difícil obtener información de los parámetros: en este proyecto se ha requerido probar una gran cantidad de modelos, utilizando como datos de entrada todas las combinaciones posibles de los tipos de variables disponibles para poder conocer qué variables influyen más para la predicción . En cambio, con el modelo lineal, se obtiene información directa de las variables, indicando de qué manera éstas son significativas, y cuánto contribuyen de media a la variable dependiente.
5. Por otra parte, en las redes neuronales es muy difícil interpretar qué está ocurriendo por dentro: cada neurona tiene sus parámetros, y seguirlos es muy difícil. Se trata casi de una caja negra que obtiene resultados. Por ello, si hay algún problema, es muy difícil saber dónde se encuentra.
6. Por último, cabe mencionar que es más costoso computacionalmente utilizar las redes neuronales. En el ordenador utilizado para realizar este proyecto, se requería como mínimo 4-5 minutos para ejecutar una red neuronal. Aunque no parezca mucho, es más tiempo de lo que parece: por ejemplo, para ejecutar los 7 modelos de redes neuronales en un segmento completo, se ha requerido como mínimo 4x7=28 minutos de ejecución. A este tiempo hay que multiplicarlo por las veces que hay que ejecutar el mismo modelo para obtener un valor medio lo más fiable posible. Además, hay que multiplicarlo por los 12 segmentos de datos en los que se ha probado las redes. Como consecuencia, se ha requerido de varias semanas para obtener los resultados con redes neuronales. En cambio, la ejecución de los modelos lineales requieren solamente unos segundos para obtener los resultados.

5. Conclusiones y trabajos futuros

La realización de este proyecto ha permitido estudiar exhaustivamente tanto los aspectos más importantes de los seguros de automóviles, como los diferentes aspectos clave en las ramas de la estadística y la inteligencia artificial. Los análisis y resultados de este estudio pueden ser de bastante utilidad tanto para las disciplinas que se dedican al análisis de datos, como para las empresas en el entorno asegurador y empresas de peritaje.

Los resultados obtenidos en este proyecto se consideran muy satisfactorios dado que se ha conseguido completar todos los objetivos planteados. En concreto, se ha completado exitosamente las siguientes tareas:

- Desarrollo de modelos predictivos del coste total, coste de las piezas, horas de mano de obra en pintura y horas de mano de obra en chapa requeridos en el automóvil tras un siniestro.
- Determinación de los datos que más influyen en la predicción de estas variables.
- Análisis exhaustivo de las ventajas y desventajas de los algoritmos utilizados.

5.1 Resumen del estudio

Antes de comenzar este proyecto, se ha tenido que realizar un análisis exhaustivo para ver en qué medida podrían ser útiles el desarrollo de los objetivos planteados para las empresas del entorno asegurador y pericial. Para ello, se ha requerido efectuar un estudio en profundidad sobre los seguros de automóviles y las disciplinas que se dedican al análisis de datos. Al existir dos disciplinas diferentes para la predicción de datos (estadística y aprendizaje automático), se decidió realizar este estudio mediante estos dos enfoques, con el objetivo de analizar las características esenciales de cada una y las diferencias entre las mismas.

Una vez confirmado los beneficios del proyecto, se procedió a estudiar la viabilidad en la implementación de los algoritmos. Para ello, se analizó los diferentes programas software y lenguajes a utilizar para desarrollar los modelos predictivos, y se concluyó que H2O era la mejor librería a utilizar, junto al lenguaje R, dado que ambos son de código abierto y disponen de gran cantidad de documentación y soporte de la comunidad . Posteriormente se examinó los datos de peritaje de siniestros disponibles, con el objetivo de obtener un análisis descriptivo y extraer así las características más importantes de las variables a predecir.

Tras todo este estudio preliminar, se procedió a realizar pruebas de rendimiento en las redes neuronales utilizando diferentes parámetros (número de capas , número de neuronas, etc), con el objetivo de obtener la mejor configuración para predecir las variables de interés. Se empleó una gran cantidad de tiempo para realizar estas pruebas y finalmente se concluyó experimentalmente que para este caso, las redes neuronales con una estructura de autoencoder de $(n, \frac{n}{2}, \frac{n}{2}, n)$ (en donde n es el número de neuronas y coincide con el número de variables de entrada), eran los que obtenían los mejores resultados.

Tras determinar la mejor estructura en la red neuronal, se procedió a ejecutar dicho modelo en todos los segmentos de datos, consiguiendo una tasa de aciertos media entre el 65-80%, pero obteniendo a la vez una gran dispersión de los errores, lo que indicaba que a pesar de que la mayoría de veces los modelos predecían correctamente, había ocasiones en los que sus predicciones se equivocaban de manera drástica. Para solucionar este problema, se decidió utilizar los modelos en un subconjunto de datos correspondiente a los tres primeros cuartiles de las variables a predecir, y con ello se consiguió reducir de manera extraordinaria la dispersión de los errores.

Respecto al estudio estadístico, primero se tuvo que realizar un estudio exhaustivo sobre qué distribución podían seguir las variables a predecir. Hubo problemas ajustando los datos a una distribución teórica, dado que al principio no se estaban utilizando las librerías adecuadas para el estudio, pero tras un análisis más detallado sobre los diferentes paquetes de R, se consiguió demostrar que estas variables se ajustaban bien a una distribución normal, por lo que se utilizó el modelo lineal para el análisis predictivo.

Al probar el modelo lineal, se comprobó que había errores de especificación, por lo que se tuvo que reducir la complejidad del problema, reduciendo el número de variables regresoras. Tras este paso, se procedió a comprobar el cumplimiento de los supuestos del modelo lineal. Estos supuestos son muy estrictos y no se suelen cumplir en la realidad, el cual ha sido el caso de este proyecto. En concreto, no se cumplió la homocedasticidad de la varianza, por lo que se tuvo que utilizar otro tipo de estimadores (errores robustos) para lidiar con el problema.

Con este enfoque estadístico, se obtuvieron peores resultados que con las redes neuronales, con una tasa de aciertos entre el 55-80%, pero por otra parte, se consiguió extraer información importante de los siniestros a nivel de variable: se consiguió saber en qué medida afectaba cada regresor a las variables explicativas, cosa que no es posible con las redes de neuronas.

A pesar de todas las dificultades por las que se ha tenido que pasar para el completo desarrollo del proyecto, se considera que los resultados obtenidos son muy satisfactorios: se han desarrollado modelos para la predicción de las variables de interés, lo que resultará bastante útil para automatizar el proceso de peritaje y luchar contra el fraude de seguros; y por otra parte, se ha llegado a obtener información bastante valiosa para las diferentes ramas de análisis de datos.

5.2 Trabajos futuros

Con este proyecto, se pueden establecer varias ideas para realizar estudios relacionados y mejorar así las herramientas utilizadas por las empresas de seguros.

Se ha podido comprobar que las redes neuronales tienen mejor capacidad predictiva que los modelos clásicos de la estadística, pero que éste último, proporciona información bastante interesante a nivel de variables predictoras. La idea consiste en utilizar conjuntamente las ventajas de ambas disciplinas: será interesante realizar primero estudios de los regresores mediante herramientas estadísticas para ver qué variables son las más significativas en la

predicción, y luego utilizar dichas variables para desarrollar los modelos de redes neuronales. Con este enfoque, se conseguirán modelos que no se han entrenado con datos poco inútiles para la predicción.

También podría resultar interesante proporcionar a los modelos predictivos los datos de una manera diferente: en vez de dividir éstos en 27 zonas diferentes, se podría utilizar datos de manera más precisa, es decir, dar directamente como variables de entrada cada pieza afectada del siniestro. Esto no serviría para los modelos lineales, dado que tendría problemas de especificación, pero podría usarse para mejorar los resultados en las redes neuronales.

Por otra parte, queda pendiente analizar de manera detallada los datos superiores al 3º cuartil, y ver qué información se puede extraer de ellas. A su vez, resultaría interesante combinar los proyectos realizados por González (2015) y Anca (2015) con este estudio. En estos dos proyectos se dividieron los datos en diferentes clústeres por severidad según su coste. Por ello, sería interesante dividir los datos en dichos clústeres, y probar los modelos de predicción en ellos, en vez de haberlos dividido en cuartiles.

Será interesante para las empresas de seguros utilizar estos modelos predictivos para realizar un programa informático que pueda automatizar a tiempo real la comparación de los informes de costes de los siniestros aportados, con las predicciones de los modelos, y ver con ello si se está cometiendo algún intento de fraude.

Por último, también será de gran utilidad cambiar los modelos de tasación a priori, acorde a la precisión de predicción en el segmento en cuestión: si es más fácil predecir, se podrán establecer precios más justos, mientras que si es más difícil de predecir el coste, se podrán establecer rangos de costes más amplios.

6. Planificación

En este apartado se detalla la planificación realizada para el desarrollo de este proyecto, tanto la planificación inicial realizada al empezar el mismo, como la planificación final adaptada a aspectos no previstos inicialmente.

En la primera planificación, se separó el proyecto en 6 fases principales. Por otra parte se asignó un 10% adicional al tiempo total del proyecto para los posibles imprevistos en la realización del mismo. Las fases son las siguientes:

- Investigación: se trata de la fase inicial de formación y documentación sobre los algoritmos que se usarán en el proyecto y del entorno en el que se engloba (15 días).
- Configuración del entorno: búsqueda de los elementos que van a abarcar el proyecto, instalación de las herramientas a utilizar y familiarización de las mismas (7 días).
- Experimentación de algoritmos: realización de las pruebas necesarias de los algoritmos para dar solución al problema planteado en el proyecto (40 días).
- Análisis de los resultados: análisis de los resultados obtenidos tanto para las pruebas con redes neuronales como las pruebas con modelos lineales generalizados (14 días).
- Elaboración de la memoria: elaboración de este documento (30 días).
- Revisión de la memoria (7 días).

La planificación inicial estimaba una duración total de 113 días, a los que se le añadió 12 días para mitigar el impacto de posibles imprevistos sumando un total de 125 días. Se asignó a cada día de trabajo 4 horas, lo que hace un total de 500 horas de trabajo.

Esta planificación tuvo que modificarse durante la realización del proyecto, ya que en un primer análisis de los resultados se descubrió que se podían mejorar los modelos, por lo que la fase de experimentación se alargó 10 días y la de análisis 5 días. Por otra parte, la elaboración de la memoria disminuyó en una semana ya que en cada fase anterior se realizó parte de la misma.

La planificación final muestra una duración total del proyecto de 128 días, con una media de 4 horas trabajadas al día, sumando un total de 512 horas de trabajo.

	Nombre de tarea	Duración	Comienzo	Fin
1	Investigación	15 días	lun 03/10/16	lun 17/10/16
2	Configuración del entorno	7 días	mar 18/10/16	lun 24/10/16
3	Experimentación de algoritmos	40 días	mar 25/10/16	sáb 03/12/16
4	Análisis de los resultados	14 días	dom 04/12/16	sáb 17/12/16
5	Elaboración de la memoria	30 días	dom 18/12/16	lun 16/01/17
6	Revisión de la memoria	7 días	mar 17/01/17	lun 23/01/17
7	Tiempo de ajuste	12 días	mar 24/01/17	sáb 04/02/17

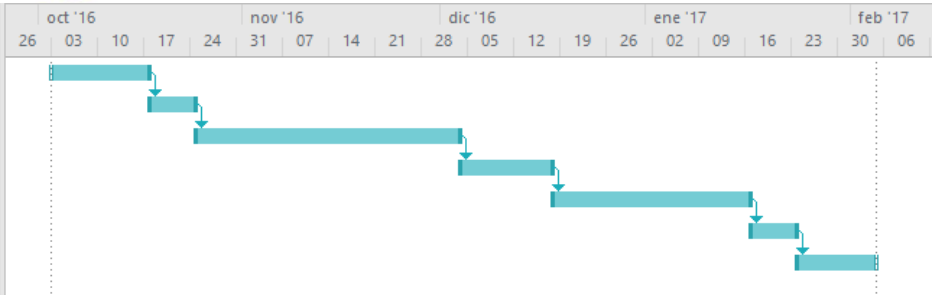


Ilustración 11: Planificación inicial del proyecto

Nombre de tarea	Duración	Comienzo	Fin
Investigación	**15 días**	**lun 03/10/16**	**lun 17/10/16**
Documentación sobre seguros y peritación	3 días	lun 03/10/16	mié 05/10/16
Documentación sobre Redes de neuronas y Modelo lineal	12 días	jue 06/10/16	lun 17/10/16
Configuración del entorno	**7 días**	**mar 18/10/16**	**lun 24/10/16**
Búsqueda y elección de herramientas	5 días	mar 18/10/16	sáb 22/10/16
Instalación	2 días	dom 23/10/16	lun 24/10/16
Experimentación de algoritmos	**62 días**	**mar 25/10/16**	**dom 25/12/16**
Primeras pruebas	8 días	mar 25/10/16	mar 01/11/16
Experimentación con redes neuronales	17 días	mié 02/11/16	vie 18/11/16
Experimentación con estadística	15 días	sáb 19/11/16	sáb 03/12/16
Mejora de resultados	10 días	vie 16/12/16	dom 25/12/16
Análisis de resultados	**29 días**	**dom 04/12/16**	**dom 01/01/17**
Primer análisis	12 días	dom 04/12/16	jue 15/12/16
Segundo análisis	7 días	lun 26/12/16	dom 01/01/17
Elaboración de la memoria	**99 días**	**mar 11/10/16**	**mar 17/01/17**
Introducción	2 días	mar 11/10/16	mié 12/10/16
Situación actual	5 días	jue 13/10/16	lun 17/10/16
Diseño de la solución	3 días	lun 02/01/17	mié 04/01/17
Evaluación de resultados	10 días	jue 05/01/17	sáb 14/01/17
Conclusiones	3 días	dom 15/01/17	mar 17/01/17
Revisión de la memoria	7 días	mié 25/01/17	mar 31/01/17

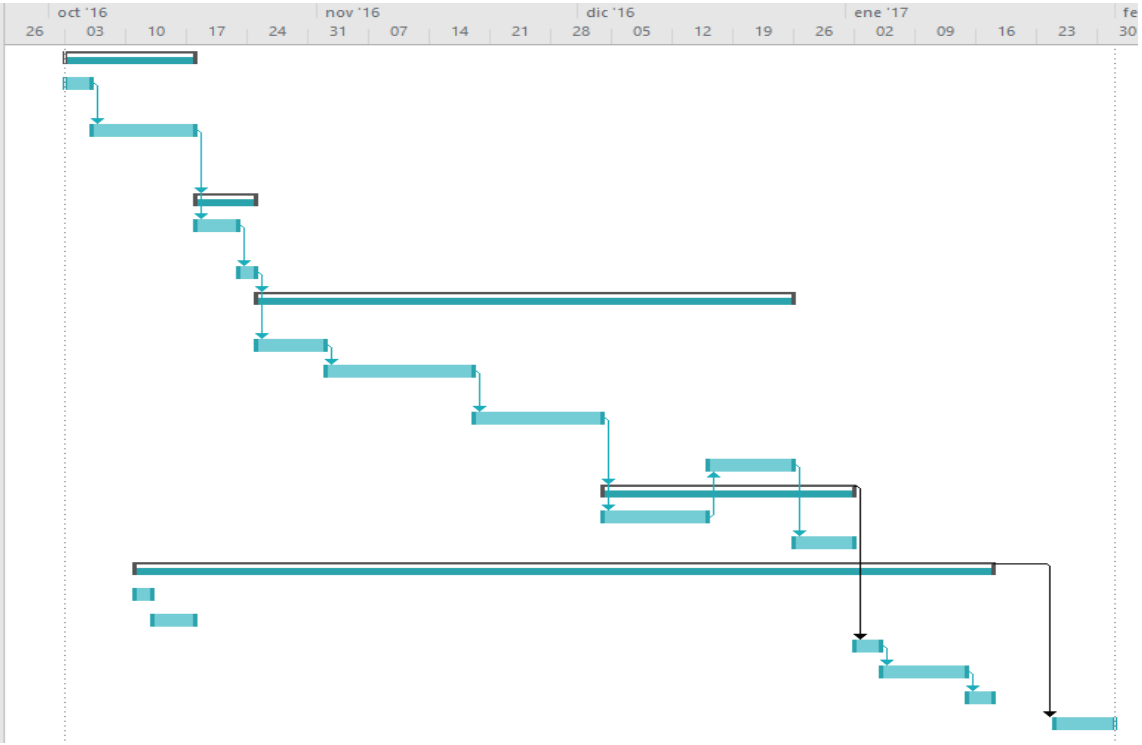


Ilustración 12: Planificación final del proyecto

7. Presupuesto

A continuación se muestra el presupuesto que se ha requerido para el desarrollo de este proyecto, detallando todos los costes incurridos.

Para el cálculo de los costes de equipamiento y software, se ha tenido en cuenta los periodos de amortización y los periodos de tiempo en los que se ha usado cada herramienta.

PERSONAL					
Nombre	Coste por hora		Horas trabajadas		Coste total
Ming-Da Liu Zhang	24,00 €		512		12.288,00 €
				Total	12.288,00 €
MATERIAL					
Descripción	Coste	Tiempo de uso	Uso en el proyecto	Periodo de amortización	Coste imputable
Acer Aspire E-571	600,00 €	4 meses	100%	48 meses	50,00 €
				Total	50,00 €
SOFTWARE					
Descripción	Coste	Tiempo de uso	Uso en el proyecto	Periodo de amortización	Coste imputable
Microsoft Windows 10	135,00 €	4 meses	100%	48 meses	11,25 €
Microsoft Office Pro 2016	539,00 €	4 meses	50,00%	48 meses	22,46 €
				Total	33,71 €

Tabla 64: Presupuesto desglosado

TOTAL	
Concepto	Coste
Personal	12.288,00 €
Material	50,00 €
Software	33,71 €
Total sin impuestos	12.371,71 €
IVA (21%)	2.598,06 €
Total	14.969,77 €

Tabla 65: Coste total del proyecto

8. Bibliografía

Anca, G. (2015). *Clasificación de patrones de siniestros aplicado técnicas de agrupación y segmentación* (Trabajo Fin de Grado, Universidad Carlos III de Madrid). Recuperado de: http://e-archivo.uc3m.es/handle/10016/23044

Belsley, D. A. (1991), *Conditioning Diagnostics*, New York: Editorial John Wiley.

Breusch, T. S., y Pagan, A. R. (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica: Journal of the Econometric Society*, 1287-1294.

Candel, A., Parmar, V., Ledell, E. y Arora, A. (2016). *Deep Learning with H20*. Recuperado de: http://docs.h2o.ai/h2o/latest-stable/h2o-docs/booklets/DeepLearningBooklet.pdf. Último acceso: 15/10/2016

Chatterjee, S., y Price, B. (1991). *Regression Diagnostics*, New York: Editorial John Wiley.

Dietterich, T. (1995). Overfitting and undercomputing in machine learning. *ACM computing surveys (CSUR)*, 27(3), 326-327.

Dirección General de Tráfico. (2015). *Anuario Estadístico General 2015*, [Informe]. Recuperado de: http://www.dgt.es/Galerias/seguridad-vial/estadisticas-e-indicadores/publicaciones/anuario-estadistico-de-general/Anuario-estadistico-general-2015.pdf. Último acceso: 10/10/2016

Dirección General de Tráfico. (2015). *La DGT pone en marcha nuevas medidas para la gestión de la velocidad*, [Nota de prensa]. Recuperado de: http://www.dgt.es/es/prensa/notas-de-prensa/2015/20150219-La-DGT-pone-en-marcha-nuevas-medidas-para-la-gestion-de-la-velocidad.shtml. Último acceso: [10/10/2016]

Dirección General de Tráfico. (2015). *Las principales cifras de la siniestralidad vial. España 2015.* [Informe]. Recuperado de: http://www.dgt.es/Galerias/seguridad-vial/estadisticas-e-indicadores/publicaciones/principales-cifras-siniestralidad/Las-principales-cifras-2015.pdf. Último acceso: [10/10/2016]

Guillén Estany, M. et al (2005). *El seguro de automóviles: estado actual y perspectiva de la técnica actuarial.* Majadahonda: Editorial MAPFRE.

Goldberger, A.S., (1991). *A Course in Econometrics*. London.

Gorunescu, F. (2011). *Data Mining - Concepts, Models and Techniques.* Craiova (Rumanía). Editorial Springer.

González, E. (2015). *Análisis de datos aplicado a siniestros de automóviles* (Trabajo fin de grado, Universidad Carlos III de Madrid). Recuperado de http://e-archivo.uc3m.es/bitstream/handle/10016/23045/TFG_Eduardo_Gonzalez_Gonzalez.pdf?sequence=1

Hand, D. Mannila, H. y Smyth, P. (2001). *Principles of Data Mining*. Massachusetts, Editorial The MIT Press.

Hecht-Nielsen, R. (1988). Theory of the backpropagation neural network. *Neural Networks*, *1*(Supplement-1), 445-448.

Hilera, J.R. y Martínez, V.J. (1995). *Redes neuronales artificiales. Fundamentos, modelos y aplicaciones*. Madrid, Editorial RA-MA.

Iturgoyen, M. (Sin fecha). *La tipología del fraude en el seguro de automóviles en España.* MAPFRE. Recuperado de: https://www.fundacionmapfre.org/documentacion/publico/i18n/catalogo_imagenes/grupo.cmd?path=1035854. Último acceso: 15/10/2016

Le, Q. V. (2015). A Tutorial on Deep Learning Part 1: Nonlinear Classifiers and The Backpropagation Algorithm.

Le, Q. V. (2015). A Tutorial on Deep Learning Part 2: Autoencoders, Convolutional Neural Networks and Recurrent Neural Networks.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436-444.

Ley 15/1999 de 13 de diciembre, de Protección de Datos de Carácter Personal (BOE núm 298, de 14 de diciembre de 1999).

Ley 08/2004 de 29 de octubre, de Ley sobre responsabilidad civil y seguro en la circulación de vehículos a motor (BOE núm 267, de 5 de noviembre de 2004).

McCullagh, P. y Nelder, J.A (1983). *Generalized Linear Models.* Londres, Editorial Chapman and Hall

Nelder, J. A y Wedderburn, R. W. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society,* 135 (3), 370-384.

Pérez Torres, J.L. (2001). *Fundamentos del seguro*. Barcelona, Editorial UMESER.

Piboongungon, T., Aalo, V. A., Iskander, C. D., & Efthymoglou, G. P. (2005). Bivariate generalised gamma distribution with arbitrary fading parameters. *Electronics Letters*, *41*(12), 709-710.

Rencher, A. C. y Schaalje, G.B. (2007). *Linear Models in Statistics.* Utah, Editorial John Wiley & Sons Inc. Publication

Rosenblatt, F. (1957). *The perceptron, a perceiving and recognizing automaton Project Para*. Cornell Aeronautical Laboratory.

Samuel, A. L. (1959). Some Studies in Machine Learning - Using the Game of Checkers. *IBM Journal of research and development, 3*(3), 210-229.

Savin, N. E, y White, K.J.(1977). The Dublin-Watson Test for Serial Correlation with Extreme Sample Sizes or Many Regressors. *Econometrica: Journal of the Econometric Society,* 1989-1996.

UNESPA. (2015). *Las estafas al seguro cuestan 550 millones de euros en 2015.* [Nota de prensa]. Recuperado de: http://www.unespa.es/adjuntos/fichero_4155_20160421.pdf. Último acceso: 10/10/2016.

Widrow, B., & Lehr, M. A. (1990). 30 years of adaptive neural networks: perceptron, madaline, and backpropagation. *Proceedings of the IEEE*, *78*(9), 1415-1442.

Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle. (2007). *Greedy layer-wise training of deep networks In Advances in Neural Information Processing Systems*.

9. Anexos

9.1 Anexo A: English summary

9.1.1 Introduction

In the current society, automobiles has become the most common way to transport peoples or goods. According to the last report of the General Direction of Traffic, in 2015 there were over 31 million of vehicles registered in Spain. However, because of this huge number of vehicles, people have to face the risks of accidents in their daily life. There are still a lot of accidents in Spain: in 2015 there was registered 97756 car accidents (DGT, 2015). This accidents generate unknown damages, and it is necessary to do an appraisal to evaluate its costs. The high risks of accidents, along with the costs that they imply, creates the necessity of a compensatory system: the automobile insurance.

The driver is the person who has to pay the damage caused by an accident. In order to cover these risks, they hire an automobile insurance. With the insurance, they just have to pay a fare, and then the insurance company becomes the responsible of the payment in case of an accident. Therefore, it is really important for the insurance companies to calculate propertly how much they have to charge their clients.

On the other hand, insurance companies keep records of thousands of appraisals in their databases. With a good management and analysis of this data, they can improve their strategy decisions. Thanks to disciplines like Statistics or Machine Learning, the study of this data is possible.

Automobile insurance companies need to develop reliable models to predict the costs of accidents and fraud cases. According to a report of UNESPA (2015), the automobile insurance industry receive more than half of the fraudulent claims (53%) of all cases of fraud attempts to insurance companies.

The growth of available data, along with the new types of insurance to satisfy custom needs of the clients, demand the creation of new techniques and tools to manage, analyze and predict its patterns.

9.1.2 Objectives

The goal of this project is to develop new tools to predict interesting variables of the automobile accidents. Specificaly, we are going to predict the following values:

- Total cost to repair completely the automobile after an accident.
- Total costs of the new pieces used to repair the automobile.
- Hours of labour needed to paint the automobile
- Hours of labour needed to prepare the car sheet.

On the other hand, a study will be carried out to analyze what kind of data is more important in order to predict these variables.

In order to achieve these goals, an automobile insurance company has shared with us a database with more than four millions of appraisals of car accidents. This data is divided in 12 segments, according to attributes like the car brand, model, size, quality, etc.

These data contains information about the variables that is going to be predicted in this project, along with the information about the number of pieces in each area of the automobile, that required to be painted, repaired or replaced.

Therefore, it is going to be used two approaches to predict these data:

- Machine learning: with the use of neural networks with a deep learning arquitecture, predictive models will be created with supervised learning.
- Statistics: with the use of generalized linear models, predictive models will be created for the same cause.

The develop of these tools will be really helpful for insurance companies: it will help to detect fraud in the automobile accident appraisals, because if the real costs differs significantly with the predicted cost, this will be a hint of a fraud attempt. On the other hand, it will allow the insurance companies to calculate in a better way the fare to charge, according to the uncertainty of the predictions in the segments.

Finally, a comparison will be made about the main characteristics of the two approaches: machine learning and statistics, in order to analyze the advantages and disadvantages of both fields.

9.1.3 Solution Design

In this section, there will be explained all the decisions made in order to develop the project. Specifically, tools used (software and hardware), and the implementation of the different algorithms.

The available data is divided in 12segment. Each file corresponds to a type of vehicle according to its brand, model and size. This data has been taken in a 5-year period and together there are 4.147.715 appraisals of car accidents.

The data is saved in CSV files, which means "comma-separated values". They are documents in a tabla structure, in which each column is separated by commas, and each column represent a variable of the automobile. Each row represents an instance of the data, and the first one indicates the name of each variable.

Each appraisal is represented by 87 attributes:

- Total cost without taxes to repair the the automobile.

- Total cost of the new pieces for the car.
- Total labour hours fot painting the automobile.
- Total labour hourse for the sheet of the car.
- Number of pieces substituted in each area.
- Number of pieces repaired in each area.
- Number of pieces painted in each area.
- Number of sequence (identification of the appraisal).
- Model of the automobile (segment)

The automobile has been divided in 27 areas:

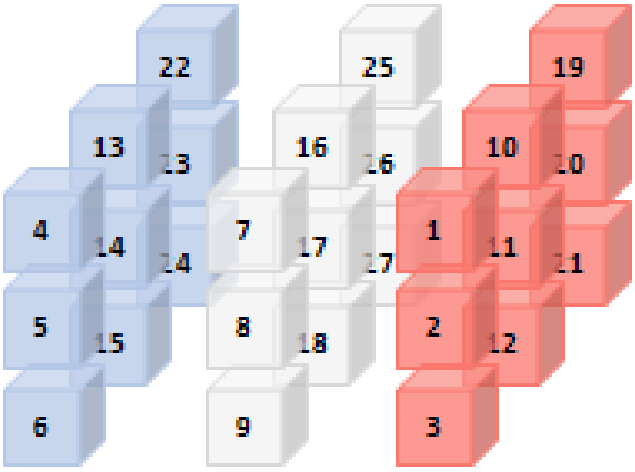


Ilustración 13: Automobile divided in 27 areas

In the following table it is shown the number of appraisals per segment

Segment	Number of appraisals
A	161698
B	712552
C	865814
D	673479
E	270407
F	42058
G	652025
H	124165
TA	215130
TB	172741
TC	171513

Tabla 66: Number of accident appraisals per segment

In this project, there has been used a multiple software tools:

- H2O: it is an oper source artifitial intelligence platform. It stands out because of the performance and scalability of its algorightms, and it is worldwide used. It offers the implementation of algorithms such as deep learning, tree decision, GLRM, etc. This package is supported in multiple languages, like R, Python, Java or Scala.
- Rstudio and R: R is one of the main languages for statistical analysis, and RStudio is a platform that helps the programmer to develop their programs.
- Other software: Microsoft Excel to create the tables and simple statistics analysis. Microsoft Word create this document
- Operating systems: Windows 10. All the software used supported this OS, so there was not any necessity to use Linux.

In order to develop the different models to predict the data, it was necessary to make some data preprocessing:

- Normalization: to deal with the huge range in the data, it is necessary to normalize the data before using it on the neural networks.
- Aleatoritation: to avoid overfitting
- Data filtering: to avoid useless data.
- Data division: in training set and validation set, in order to check the performance of the models.

9.1.4 Neural network analysis

Seven models have been tested in order to predict the different variables, according to the data provided to the neural networks: substitution, painting, and reparation.

In the following table, it is summarized the best model for the prediction of each variable:

Dependend variable	Best model
Total cost	Painting, Substitution
Cost of pieces	Painting, Substitution
Labour in painting	Painting, Reparation, Substitution
Labour in sheet	Painting, Reparation, Substitution

Tabla 67: Best models in each variable

In the following table, it is summarized the performance of the neural networks:

Used data	Descriptives	Total Cost	Cost of pieces	Labour in paintng	Labour in sheet
Todos los datos	Mean error	299,55 €	206,34 €	1,1 horas	1,83 horas
	Standard deviation	458,18 €	384,78 €	1,62 horas	2,61 horas
	Performance	73,02%	65,41%	81,51%	65,81%
0-3º cuartil	Mean error	131,43 €	74,82 €	0,67 horas	0,85 horas
	Standard deviation	116,87 €	85,38 €	0,58 horas	0,72 horas
	Performance	74,18%	68,00%	81,58%	71,38%

Tabla 68: Summary of results with neural networks

The best variables to predict are the labour in painting, with a performance of 81%, and a mean error of 1,1 hours per prediction if we use all the data. The same performance is obtained if we use the data from 0 to the third quartile, but the mean error is reduced to 0,67 hours. The second best variable to predict is the total cost, with a performance of 73 and 74%, and mean errors of 300€ and 117€.

The labour in sheet is placed in the third place, with a mean error of 1,83 hours and 0,85 hours, and performances of 65,81% and 71,38%, in all the data and the subset of 0-3º quartile respectively.

We get the worst result in the prediction of the cost of pieces, with a performance of 65% using all the data, and 68% using the subset of data from 0 to the third quartile. Anyway, the prediction of this variable is the least important, because the insurance companies care mor about the prediction of the total cost, which has better performance.

If we use all the data in the neural networks, in all cases the standard deviations are bigger than the mean error obtained. This dispersion of the error occurs because the neural networks do not have the capacity to predict extremely high values. The bigger the value, the worse the prediction.

But we can see that if we use the subset from 0 to the third quartile, the standard deviation decreases a lot. This makes easier the interpretation of the results, and makes the prediction more precise.

Finally, if we rank the prediction per segments, generally segments A, B and C obtain the best predictions, while F and H the worst predictions.

9.1.5 Linear model analysis

In order to develop the linear models, it was necessary to check that the models met all the assumptions: linearity, homocedasticity and multicollinearity. In all the models, the assumption of homocedasticity was not met, so it was necessary to use robust standard errors in order to create a correct model. Furthermore, it was not possible to develop the models using all the variables, because they were producing negative parameters, which makes no sense for interpretation. Because of that, it was necessary, all the regressors were reduced to 9, except for the prediction of the cost of the pieces, which was reduced to 3.

The following table summarizes the results obtained in the prediction of the variables.

Coefficients	Total cost	Cost of pieces	Labour in painting	Labour in sheet
Intercept	58,18	12,06	1,41	0,5
s1	139,81	112,66	0,02	0,71
s2	213,13		0,15	1,34
s3	129,25		0,04	0,93
p1	48,16	7,35	0,81	0,021
p2	22,07		0,66	0,019
p3	64,99		0,57	0,32
r1	12,4	2,73	0,21	0,69
r2	24,69		0,43	0,55
r3	15,97		0,43	0,75
Mean error	203,05	145,77 €	0,98	2,02
Standard deviation	271,05	224 €	1,10	2,73
Performance	71,64%	61,49%	78,58%	58,42%

Tabla 69: Summary of results with the linear model

Overall, the results are good in the prediction of the total cost, with a performance of 72%, and the prediction of labour in painting, with a performance of 78%. The prediction in the cost of pieces and labour in sheet, just achieved a performance of 61% and 58% respectively, so they are considered mediocre results.

Analyzing the βs, we can see in the table that for the estimation of the total cost and the cost of pieces, the most significant variable is the substitution of pieces, followed by

the painting, and staying in the last position the reparation. In the estimation of the labour in shet, the substitution was also the most improtant attribute, but it was followed by reparation instead of painting. This last variable barely have any impact in the prediction.

Finally, in the prediction of the labour in painting, the most important variable is obviously the painting, and it is followed by reparation and substitution is in the last place.

9.1.6 Comparison between linear models and neural networks

The next table summarizes the performance obtained by both algorithms in the different regression models:

Dependent variable	Used data	LM	Neural networks
Total Cost	All	71,92%	73,02%
	0-3º quartile	70,55%	74,18%
Cost of pieces	All	56,50%	65,41%
	0-3º quartile	55,82%	68,00%
Labour in painting	All	78,41%	81,51%
	0-3º quartile	78,18%	81,58%
Labour in sheet	All	61,16%	65,81%
	0-3º quartile	68,19%	71,38%

Tabla 70: Comparison between linear model and neural networks

We can see that generally the models using neural networks obtains better performance in its predictions. In the total cost, labour in painting and labour in sheet its performance is better by 3%. But the biggest difference is obtained in the prediction of the cost of pieces, Here, the prediction using neural networks is 10-12% better than the predictions using linear models. This is an indication that the neural network has learnt a hidden pattern in the data, that was not possible to detect in the linear model. Because of that, it is understood that the neural networks have more capacity to solve more complex problems.

There are some differences between linear models and neural networks:

7. In the statistics models, it is required to study previously the data before using any model: it is necessary to know what distribution of probability has the data. If the distribution is unknown, the results are really poor. This pre-analysis is not necessary in the neural networks.
8. The linear models requires that some assumptions are met, in order to have a correct specification of the model. If they are not met, the results are poor. On the other hand, the neural networks do not require any kind of assumptions, so they are easier to develop.

9. The develop of linear models is less automatic: after creating the models, it is necessary to check step by step that all the assumptions are met. If they are not met, it is required to find a way to fix them. If they are not fixed, the obtained results are not possible to be analyzed.
10. It is not all disadvantages in the linear models. They offer a different approach, and put a lot of effort inferring the different parameters of the model, while the neural networks just focus on the prediction performance. In the neural networks, it is a lot more difficult to obtain individual information about the regressors. In this project, it was necessary to develop 7 models to obtain some information about the individual predictors. In contrast, with the linear model you can obtain this information directly, and they provide the signifcance of each variable, and how much they contribute on average to the final prediction.
11. On the other hand, it is really difficult to interpret what is happening inside the neural network: each neuron has its own parameters, and following them is an impossible task. So if there is a problem with the neural network, it is really difficutl to know what is happening.
12. Finally, we have to mention the computational cost of each algorithm. Neural networks require a lot more computation. In the computer used to develop this project, the neural networks required about 4-5 minutes to be executed in a segment. It looks like it is not that much time, but for example, in order to execute the 7 neural networks in a segment, it was required as minimum 4x7=28 minutes. And we have to multilpy it for the times you execute those neural networks, in order to get a mean performance in the segment. Furthermore, this was executed in the 12 segments of data. As consequence, it required some weeks to test the neural networks and obtain its results. On the other hand, the execution of linear models are instant, and they just require some seconds to obtain the results.

9.1.7 Conclusions

The development of this project has allowed to study thoroughly the most important characteristics in the insurance environment, along with the key aspects in the fields of statistics and artifitial intelligence. The analysis and results in this study can be really useful for both the disciplines that look for the data analysis, and the insurance companies.

The results obtained in this project are more than satisfactory because all the objectives were met. Specifically:

- The development of predictive models for the total cost, cost of pieces, labour in painting and labour in sheet required in an automobile after suffering an accident.
- Analyze the most important data for the prediction of these variables.

- Analysis of the advantages and disadvantages of the algorithms used in the project.

9.1.7.1 Summary of the development of this project

Before starting the project, it was necessary to carry out an exhaustive analysis, in order to know if the results of this project would be useful to the automobile insurance companies. Because of that, it was required to do an in-depth study about the automobile insurance, and the different disciplines for the data analysis.

There are two different disciplines for the prediction of the data (statistics and machine learning), so because of that, it was decided that it would be interesting develop this study using these two different approaches, in order to analyze the characteristics and the differences of these fields.

Once the benefits of this project were known, it was time to study the viability of the implementation of the proposed algorithms. Because of that, it was necessary to analyze all the possible software programs and programming languages in order to develop these predictive models. The conclusion of this analysis was that the best choice was to use the H2O library, along with the programming language R, because both of them are open source and have a lot of documentation and community support. After that, it was necessary to study the available database, in order to obtain the main descriptive statistics of the variables that were going to be predicted.

After this preliminary analysis, an in depth experimentation was carried out with neural networks, using different parameters (number of layers, number of neurons, etc.), in order to determine the best configuration to predict the different variables. A lot of time was spent in this phase, and at the end, the experimental conclusion was that the neural networks with an autoencoder structure $(n, \frac{n}{2}, \frac{n}{2}, n)$ (where n is the number of neurons, which is the same number as the number of input variables), were the neural networks that obtained the best results.

After knowing the best structure for the neural network, this model was executed in all the datasets, and they achieved a performance between 65-80%, but with big dispersion of the errors, which meant that despite having accurate predictions most of the times, sometimes these predictions were completely wrong. In order to solve this problem, these models were executed in a subset of the data, from the minimum to the third quartil. With the reduction of the data, a big reduction of the dispersion of errors were achieved.

Regarding the statistical study, firstly it was necessary to carry out an exhaustive study about what kind of distributions were the most fittable in the dependent variables. There were multiple problems fitting this variables to a theoretical distribution, because at the beginning, the wrong library was being used. After a more detailed analysis of the different libraries in R, it was proved that these variables fitted correctly a normal

distribution. Because of that, it was possible to use the linear models in order to create the predictive models.

After trying out the linear models, it was discovered that there were specification problems, so it was necessary to carry out a reduction of the problem, by reducing the number of predictors. After this step, it was necessary to study the fulfillment of the assumptions of the linear model. These assumptions are very strict, and they are not usually met in real world problems. This was the case in this project. Specifically, the homocedasticity assumption was not met, so it was necessary to use robust standard errors in order to solve this problem.

The statistical approach achieved worse results than the neural networks, with performance of 55-80%. On the other hand, with this study it was possible to obtain important information about the car accident in a variable level. They showed what regressors were the most important for the prediction of the variables, and this is not possible to get using neural networks

Despite all the difficulties that were faced in this project, it is considered that the obtained results are really satisfatory: there were developed different models for the prediction of these variables, and they will be really useful to automate the process of detecting the fraud attempts. On the other hand, some interesting information was concluding about the disciplines of machine learning and statistics.

9.1.7.2 Future research

Thanks to this project, some ideas of related research can arise in order to improve the obtained results and thus improve the tools used by the insurance companies.

It was possible to prove that the neural networks have better predictive performance than the typical linear models. But hte linear models provides interesting information about the predictors. This idea consists in using the advantages of both fields: it will be interesting to analyze firstly the data with statistical tools, in order to learn what variables are the most significative for the prediction, and then using these variables to develop the neural networks. With this approach, it is possible to obtain neural networks that have not be trained with useless data.

It would be also interesting to use another approach of the dataset: instead of dividing the data in 27 different areas, it would be interesting to use the data in a more detailed way. That is to say, to give as input variables every piece affected in the automobile. This approach would not be useful for the linear models, because fot he specification problems, but it could be used to improve the results in the neural networks.

On the other hand, it remains pending the analysis of the data over the third quartile, and study what kind of information can be get with this. At the same time, it would be interesting to combine the proyects developed by González (2015) and Anca (2015)

with this study. In these two projects, they divided the data in different clusters, according to the severity of the accident by cost. Thus, it would be really interesting to divide the data in those clusters, and try there the predictive models developed in this project.

It would be also really benefitial for the insurance companies to use the predictive models of this project in order to develop a computer program, in order to automate in real time the comparison between the cost reports of the accidents, and the predictions of the models, and check the discrepancy between them in order to detect any fraud attempt.

Finally, it will be really useful to change the fare models, according to the performance obtained in each segment: if it is easier to predict in a segment, this can be translated to fairer fares, while if it is more difficult to predict the cost, widen the rank of these fares.

9.2 Anexo B: Resultados del modelo lineal

Coeficientes	Estimación	Std. Error	Estadístico t	Pr(>|t|)
Intercepto	76,17	1,54	49,41	<2e-16
SZ1	43,45	18,78	2,31	0,02
SZ2	-111,88	1,17	-95,64	<2e-16
SZ3	391,82	11,10	35,29	<2e-16
SZ5	107,13	1,14	93,61	<2e-16
SZ6	391,27	10,71	36,52	<2e-16
SZ7	120,38	1,78	67,59	<2e-16
SZ8	-149,66	0,78	-193,01	<2e-16
SZ9	66,30	4,71	14,07	<2e-16
SZ10	88,11	3,29	26,82	<2e-16
SZ11	-197,48	2,01	-98,19	<2e-16
SZ12	397,30	11,92	33,33	<2e-16
SZ13	92,42	3,38	27,34	<2e-16
SZ14	216,77	2,14	101,10	<2e-16
SZ15	356,20	11,05	32,23	<2e-16
SZ16	313,80	5,36	58,55	<2e-16
SZ17	431,89	3,09	139,57	<2e-16
SZ18	199,92	15,25	13,11	<2e-16
SZ19	265,15	11,58	22,91	<2e-16
SZ20	-116,38	1,73	-67,32	<2e-16
SZ21	463,79	35,91	12,92	<2e-16
SZ22	222,65	11,82	18,84	<2e-16
SZ23	119,49	1,88	63,48	<2e-16
SZ24	396,06	39,86	9,94	<2e-16
SZ25	142,88	2,90	49,21	<2e-16
SZ26	-96,80	1,22	-79,08	<2e-16
SZ27	494,68	6,54	75,66	<2e-16
PZ2	25,39	3,40	7,47	8,32e-14
PZ3	-348,31	57,57	-6,05	1,45e-09
PZ5	33,30	3,53	9,43	<2e-16
PZ6	-337,55	57,62	-5,86	4,69e-09
PZ8	69,34	1,82	38,18	<2e-16
PZ11	43,50	2,50	17,42	<2e-16
PZ14	31,92	2,52	12,68	<2e-16
PZ16	40,89	6,92	5,91	3,40e-09
PZ20	53,17	3,42	15,54	<2e-16
PZ23	60,03	3,47	17,31	<2e-16
PZ25	-67,53	17,30	-3,90	9,49e-05
PZ26	100,41	2,33	43,14	<2e-16
PZ27	-122,74	9,24	-13,28	<2e-16

RZ2	30,59	3,90	7,84	4,68e-15
RZ5	25,74	4,05	6,36	1,99e-10
RZ7	-341,55	21,96	-15,56	<2e-16
RZ12	79,96	11,78	6,79	1,15e-11
RZ13	85,83	23,13	3,71	0,000207
RZ14	13,04	3,24	4,02	5,88e-05
RZ15	76,41	10,71	7,13	9,96e-13
RZ16	101,51	8,70	11,67	<2e-16
RZ17	-311,71	21,54	-14,47	<2e-16
RZ20	51,88	4,04	12,86	<2e-16
RZ23	52,07	4,09	12,73	<2e-16
RZ25	184,56	73,62	2,51	0,012181
RZ26	-23,95	2,42	-9,91	<2e-16
RZ27	92,40	15,14	6,10	1,06e-09

Tabla 71: Modelo lineal segmento A en el coste total con todas las variables significativas

Coeficientes	Estimación	Std,Error	Estadístico t	Pr(>|t|)
Intercepto	12,90	1,34	9,63	<2e-16
SZ1	30,09	14,64	2,06	0,04
SZ2	-85,09	0,92	-92,33	<2e-16
SZ3	295,52	8,53	34,66	<2e-16
SZ4	-30,01	15,23	-1,97	0,04
SZ5	80,08	0,90	89,05	<2e-16
SZ6	325,63	8,20	39,72	<2e-16
SZ7	105,30	1,39	75,93	<2e-16
SZ8	119,79	0,61	196,75	<2e-16
SZ9	56,19	3,62	15,54	<2e-16
SZ10	79,69	2,57	31,03	<2e-16
SZ11	128,66	1,62	79,47	<2e-16
SZ12	235,18	9,35	25,15	<2e-16
SZ13	-80,88	2,65	-30,52	<2e-16
SZ14	142,58	1,74	82,03	<2e-16
SZ15	-184,50	8,74	-21,11	<2e-16
SZ16	198,61	4,13	48,10	<2e-16
SZ17	379,78	2,38	159,39	<2e-16
SZ18	110,37	11,77	9,38	<2e-16
SZ19	163,50	8,97	18,24	<2e-16
SZ20	-79,14	1,38	-57,38	<2e-16
SZ21	351,45	27,39	12,83	<2e-16
SZ22	141,67	9,32	15,20	<2e-16
SZ23	-80,30	1,47	-54,49	<2e-16
SZ24	211,64	32,03	6,61	3,91e-11
SZ25	114,10	2,25	50,75	<2e-16

SZ26	63,98	1,00	64,05	<2e-16
SZ27	382,74	5,04	75,96	<2e-16
PZ2	-24,37	2,83	-8,61	<2e-16
PZ3	-276,59	43,57	-6,35	2,18e-10
PZ5	-14,78	2,91	-5,08	3,71e-07
PZ6	-180,75	45,34	-3,99	6,71e-05
PZ8	12,31	1,47	8,38	<2e-16
PZ9	-20,19	9,14	-2,21	0,027244
PZ11	-10,92	2,12	-5,14	2,76e-07
PZ12	-19,21	7,44	-2,58	0,009786
PZ13	-13,65	4,86	-2,81	0,00498
PZ14	-16,64	2,14	-7,76	8,69e-15
PZ16	-62,08	6,09	-10,19	<2e-16
PZ25	-36,68	13,69	-2,68	0,007379
PZ26	41,67	1,94	21,45	<2e-16
PZ27	-125,81	7,57	-16,61	<2e-16
RZ7	-105,42	18,12	-5,82	6,02e-09
RZ8	-36,95	1,90	-19,44	<2e-16
RZ11	-32,30	2,70	-11,95	<2e-16
RZ12	32,41	9,65	3,36	0,00078
RZ13	55,98	18,63	3,01	0,002659
RZ14	-29,67	2,73	-10,86	<2e-16
RZ16	39,94	7,35	5,44	5,43e-08
RZ17	-140,40	17,65	-7,95	1,84E-15
RZ18	-72,28	20,24	-3,57	0,000355
RZ20	-13,72	3,41	-4,03	5,70e-05
RZ23	-16,28	3,46	-4,70	2,61e-06
RZ26	-55,95	1,97	-28,33	<2e-16
RZ27	72,03	12,15	5,93	3,08e-09

Tabla 72: Modelo lineal segmento A en el coste de piezas con todas las variables

Coeficientes	Estimación	Std,Error	Estadístico t	Pr(>|t|)
(Intercept)	-6,31	1,86	-3,393	<2e-16
s1	110,60	0,68	161,95	<2e-16
s2	157,81	2,33	67,50	<2e-16
s3	-91,30	0,93	-91,02	<2e-16
p1	-6,03	1,44	-4,15	<2e-16
p2	-27,37	1,03	-26,21	<2e-16
p3	11,28	1,34	8,39	<2e-16
r1	-21,35	1,61	-13,35	<2e-16
r2	-17,97	1,55	-11,63	<2e-16
r3	-33,64	1,54	-21,77	<2e-16

Tabla 73: Modelo lineal segmento A en el coste de piezas con 9 variables

Coeficientes	Estimación	Std. Error	Estadístico t	Pr(>|t|)
(Intercept)	1,42	0,01	187,422	<2e-16
SZ1	0,29	0,10	2,892	0,003826
SZ2	0,06	0,01	11,098	<2e-16
SZ3	-0,17	0,05	-3,369	0,000755
SZ5	0,04	0,00	7,287	3,18e-13
SZ6	-0,20	0,05	-4,162	3,15e-05
SZ7	0,12	0,02	6,605	4,01e-11
SZ8	-0,05	0,00	-15,404	<2e-16
SZ9	-0,28	0,02	-13,219	<2e-16
SZ11	0,19	0,01	21,103	<2e-16
SZ12	0,15	0,05	2,887	0,003885
SZ14	0,20	0,01	21,720	<2e-16
SZ15	0,34	0,05	6,901	5,19e-12
SZ16	0,16	0,03	6,170	6,87e-10
SZ19	0,46	0,07	7,074	1,52e-12
SZ21	0,66	0,15	4,244	2,20e-05
SZ22	0,35	0,07	5,133	2,86e-07
SZ23	0,02	0,01	2,802	5,08e-03
SZ25	0,26	0,02	13,565	<2e-16
SZ27	-0,22	0,03	-7,154	8,46e-13
PZ2	0,71	0,01	48,451	<2e-16
PZ3	-0,73	0,23	-3,182	1,46e-03
PZ5	0,75	0,02	49,397	<2e-16
PZ6	-0,59	0,24	-2,450	1,43e-02
PZ8	0,89	0,01	112,039	<2e-16
PZ9	-0,95	0,05	-18,667	<2e-16
PZ10	0,19	0,03	7,315	2,60e-13
PZ11	0,74	0,01	69,051	<2e-16
PZ12	0,24	0,04	6,352	2,14e-10
PZ13	0,21	0,02	8,246	<2e-16
PZ14	0,71	0,01	65,382	<2e-16
PZ15	0,07	0,04	1,978	0,047943
PZ16	2,39	0,03	81,194	<2e-16
PZ20	0,53	0,01	35,843	<2e-16
PZ23	0,50	0,01	33,544	<2e-16
PZ26	0,75	0,01	74,190	<2e-16
PZ27	-0,47	0,04	-11,410	<2e-16
RZ2	0,16	0,02	9,508	<2e-16
RZ5	0,12	0,02	6,937	4,03e-12
RZ7	-0,76	0,11	-7,167	7,74e-13
RZ8	0,19	0,01	18,713	<2e-16
RZ11	0,34	0,01	24,546	<2e-16

RZ12	0,36	0,05	7,350	2,00e-13
RZ13	-0,30	0,10	-2,972	0,002961
RZ14	0,38	0,01	27,791	<2e-16
RZ15	0,53	0,05	11,771	<2e-16
RZ17	-0,84	0,10	-8,444	<2e-16
RZ20	0,63	0,02	36,943	<2e-16
RZ23	0,75	0,02	42,726	<2e-16
RZ25	1,00	0,31	3,251	0,001151
RZ26	0,18	0,01	17,089	<2e-16
RZ27	-0,26	0,07	-3,979	6,92e-05

Tabla 74: Modelo lineal segmento A de las horas de pintura con todas las variables

Coeficientes	Estimación	Std. Error	Estadístico t	Pr(>|t|)
(Intercept)	0,45	0,02	29,68	<2e-16
SZ1	-0,56	0,18	-3,04	0,002409
SZ2	0,63	0,01	56,04	<2e-16
SZ3	2,40	0,11	22,60	<2e-16
SZ4	0,52	0,19	2,70	0,006865
SZ5	-0,61	-0,01	55,29	<2e-16
SZ6	1,86	0,10	18,14	<2e-16
SZ7	0,86	0,02	50,21	<2e-16
SZ8	0,73	0,01	97,57	<2e-16
SZ9	0,16	0,04	3,51	0,000454
SZ10	-0,45	0,03	-14,25	<2e-16
SZ11	1,50	0,02	77,69	<2e-16
SZ12	2,95	0,11	25,69	<2e-16
SZ13	0,47	0,03	14,31	<2e-16
SZ14	1,68	0,02	81,67	<2e-16
SZ15	2,93	0,11	27,51	<2e-16
SZ16	1,56	0,05	30,54	<2e-16
SZ17	0,90	0,03	30,63	<2e-16
SZ18	2,48	0,15	16,87	<2e-16
SZ19	2,64	0,12	22,70	<2e-16
SZ20	0,98	0,02	58,74	<2e-16
SZ21	2,38	0,35	6,88	6,07e-12
SZ22	2,20	0,12	18,80	<2e-16
SZ23	0,98	0,02	54,88	<2e-16
SZ25	1,02	0,03	35,85	<2e-16
SZ26	0,63	0,01	52,99	<2e-16
SZ27	2,56	0,06	40,72	<2e-16
PZ3	-3,35	0,55	-6,09	1,15e-09
PZ6	-1,94	0,55	-3,49	0,000479
PZ8	0,24	0,02	13,67	<2e-16

PZ9	-1,14	0,11	-10,45	<2e-16
PZ10	-0,65	0,06	-11,63	<2e-16
PZ11	0,26	0,02	10,73	<2e-16
PZ12	-0,10	0,09	-1,11	0,265632
PZ13	-0,69	0,06	-12,40	<2e-16
PZ14	0,10	0,02	4,30	1,69e-05
PZ16	-0,15	0,07	-2,18	0,029272
PZ19	-1,64	1,38	-1,19	0,235126
PZ20	0,26	0,03	7,97	1,59e-15
PZ23	0,41	0,03	12,12	<2e-16
PZ25	-0,94	0,16	-5,73	1,03e-08
PZ26	0,63	0,02	27,98	<2e-16
PZ27	-0,74	0,09	-8,24	<2e-16
RZ2	0,56	0,04	15,06	<2e-16
RZ5	0,64	0,04	16,55	<2e-16
RZ7	-4,57	0,22	-20,95	<2e-16
RZ8	0,76	0,02	33,69	<2e-16
RZ9	0,81	0,33	2,49	0,012932
RZ10	1,09	0,26	4,28	1,87e-05
RZ11	0,26	0,03	8,41	<2e-16
RZ12	1,36	0,11	12,07	<2e-16
RZ13	0,69	0,22	3,11	0,001875
RZ14	-0,44	0,03	-14,27	<2e-16
RZ15	1,40	0,10	13,69	<2e-16
RZ16	1,40	0,08	16,72	<2e-16
RZ17	-3,57	0,20	-17,63	<2e-16
RZ18	1,92	0,25	7,74	1,01e-14
RZ19	4,45	1,45	3,07	0,002153
RZ20	0,89	0,04	23,04	<2e-16
RZ22	-0,95	1,35	-0,70	0,483963
RZ23	0,98	0,04	24,84	<2e-16
RZ25	2,58	0,68	3,81	0,000141
RZ26	0,53	0,02	23,08	<2e-16
RZ27	1,74	0,14	12,17	<2e-16

Tabla 75: Modelo lineal segmento A en las horas de chapa con todas las variables

Coeficientes	B	C	D	E	F	G	H	S	TA	TB	TC
(Intercept)	80,13	78,92	101,52	126,06	269,15	123,25	172,34	143,63	125,89	146,95	177,78
s1	147,53	165,59	184,81	246,65	306,50	172,40	206,15	225,26	196,69	220,30	262,45
s2	208,87	236,66	228,41	308,99	348,98	233,04	260,46	371,57	280,31	356,44	310,74
s3	121,42	131,34	149,37	208,56	277,22	151,14	175,23	234,41	175,55	203,78	220,46
p1	38,19	42,83	43,96	63,68	34,02	30,15	51,74	65,73	56,68	86,47	69,34
p2	15,42	6,88	9,92	2,62	29,13	16,91	26,60	4,18	13,22	8,49	21,50
p3	91,00	96,91	94,52	60,35	5,81	58,74	20,87	71,39	48,35	51,21	20,80
r1	1,34	6,15	3,96	24,12	4,69	17,80	11,32	10,92	3,36	11,11	34,29
r2	25,91	50,14	34,28	70,85	45,99	43,52	49,13	88,97	69,81	64,25	50,80
r3	13,66	21,01	-8,06	17,86	42,24	9,37	10,75	1,77	29,93	24,20	24,67

Tabla 76: Modelos lineales en el coste total

Coeficientes	B	C	D	E	F	G	H	S	TA	TB	TC
(Intercept)	1,13	11,26	22,55	27,71	145,04	25,36	66,26	51,54	33,57	61,50	59,15
Sustitucion	120,35	141,24	151,76	207,47	259,27	140,42	169,38	199,09	171,39	194,82	226,11
Pintura	31,52	31,36	35,46	35,66	38,40	26,68	31,67	42,62	24,39	22,49	42,74
Reparacion	14,81	17,10	20,88	28,09	45,69	18,69	22,45	30,21	18,52	18,99	31,97

Tabla 77: Modelos lineales en el coste de piezas

Coeficientes	B	C	D	E	F	G	H	S	TA	TB	TC
(Intercept)	1,31	1,36	1,39	1,53	1,61	1,55	1,70	1,45	1,58	1,61	1,43
s1	0,02	0,01	0,02	0,03	0,00	0,04	0,03	0,02	0,02	0,01	0,00
s2	0,22	0,14	0,14	0,08	0,19	0,19	0,12	0,08	0,20	0,15	0,14
s3	0,02	0,02	0,06	0,07	0,08	0,09	0,19	0,05	0,07	0,06	0,06
p1	1,09	1,04	1,01	0,99	0,98	1,02	1,02	1,01	0,97	1,01	1,09
p2	0,46	0,47	0,53	0,55	0,60	0,43	0,58	0,56	0,54	0,58	0,54
p3	1,04	1,01	0,97	0,91	0,85	0,76	0,56	1,07	0,81	0,92	0,96
r1	0,08	0,15	0,17	0,20	0,23	0,17	0,21	0,19	0,11	0,03	0,10
r2	0,51	0,54	0,47	0,50	0,42	0,56	0,54	0,51	0,53	0,43	0,46
r3	0,13	0,12	0,18	0,27	0,39	0,33	0,43	0,17	0,27	0,22	0,18

Tabla 78: Modelos lineales en las horas de pintura

Coeficientes	B	C	D	E	F	G	H	S	TA	TB	TC
(Intercept)	0,71	0,78	0,87	1,02	1,36	0,86	1,02	0,86	0,75	0,95	1,00
s1	0,63	0,57	0,60	0,70	0,72	0,67	0,67	0,74	0,61	0,66	0,67
s2	1,23	1,19	1,12	1,32	1,24	1,25	1,24	1,45	1,19	1,31	1,00
s3	0,92	0,80	0,84	1,12	1,01	0,84	0,83	1,15	0,80	0,72	0,72
p1	0,06	0,01	0,04	0,14	0,10	0,01	0,00	0,16	0,13	0,18	0,14
p2	0,01	0,01	0,03	0,06	0,13	0,03	0,07	0,18	0,09	0,06	0,10
p3	0,44	0,35	0,44	0,25	0,24	0,26	0,09	0,58	0,09	0,21	0,11
r1	0,52	0,57	0,61	0,57	0,74	0,64	0,72	0,71	0,70	0,71	0,66
r2	0,65	0,84	0,71	0,89	0,73	0,69	0,97	1,21	0,86	0,78	0,82
r3	0,41	0,42	0,44	0,59	0,67	0,64	0,61	0,47	0,82	0,72	0,68

Tabla 79: Modelos lineales en las horas de chapa

Coeficientes	B	C	D	E	F	G	H	S	TA	TB	TC
(Intercept)	154,96	176,15	200,31	244,03	352,34	193,15	243,61	230,90	217,54	259,77	297,80
s1	77,76	86,46	88,86	106,52	131,93	95,23	104,12	106,28	92,24	95,73	122,36
s2	75,83	77,89	85,65	104,04	124,55	97,31	88,49	101,39	94,99	99,54	114,16
s3	56,05	61,72	75,69	88,86	108,54	74,25	77,07	85,54	87,20	92,77	96,51
p1	57,85	52,78	49,72	52,31	31,67	45,12	44,50	61,95	47,87	46,70	52,94
p2	27,79	24,03	22,26	22,61	30,21	23,93	26,73	18,16	25,02	24,78	22,52
p3	58,90	62,04	61,65	45,60	33,07	48,54	31,91	59,24	43,65	47,08	48,05
r1	6,44	4,08	13,87	14,09	7,29	17,69	10,58	15,50	18,78	17,44	2,43
r2	29,19	45,32	37,18	44,08	33,17	41,85	53,58	59,25	52,81	45,24	45,29
r3	26,20	9,86	13,79	25,40	32,94	32,32	26,40	27,57	27,73	14,38	9,14

Tabla 80: Modelos lineales para el coste total para el 0-3º cuartil

Coeficientes	B	C	D	E	F	G	H	S	TA	TB	TC
(Intercept)	59,43	65,94	75,23	89,26	130,56	70,69	78,22	73,02	83,26	96,68	122,35
Sustitucion	42,03	47,96	51,02	64,80	79,27	50,12	54,92	62,32	62,88	66,97	72,52
Pintura	8,05	10,71	8,03	7,97	10,17	5,81	7,42	8,37	6,58	7,32	10,94
Reparacion	5,41	5,09	6,60	4,89	8,07	6,58	5,02	8,08	6,13	6,73	8,20

Tabla 81: Modelos lineales para el coste de piezas en 0-3º cuartil

Coeficientes	B	C	D	E	F	G	H	S	TA	TB	TC
(Intercept)	1,59	1,64	1,70	1,89	2,12	1,81	2,10	1,74	1,84	1,96	1,78
s1	0,01	0,02	0,03	0,03	0,01	0,01	0,02	0,02	0,00	0,01	0,01
s2	0,15	0,09	0,13	0,08	0,06	0,13	0,10	0,06	0,12	0,10	0,09
s3	0,01	0,00	0,02	0,03	0,03	0,07	0,09	0,00	0,05	0,05	0,03
p1	0,75	0,66	0,62	0,60	0,55	0,58	0,52	0,66	0,59	0,55	0,66
p2	0,43	0,38	0,36	0,36	0,35	0,37	0,37	0,41	0,39	0,41	0,39
p3	0,74	0,67	0,64	0,51	0,44	0,52	0,43	0,69	0,54	0,53	0,63
r1	0,14	0,22	0,28	0,29	0,26	0,24	0,23	0,28	0,22	0,21	0,23
r2	0,41	0,50	0,47	0,48	0,45	0,48	0,50	0,51	0,48	0,41	0,48
r3	0,19	0,25	0,30	0,40	0,43	0,36	0,36	0,29	0,30	0,25	0,24

Tabla 82: Modelos lineales para las horas de pintura en 0-3º cuartil

Coeficientes	B	C	D	E	F	G	H	S	TA	TB	TC
(Intercept)	1,37	1,39	1,48	1,71	1,95	1,41	1,49	1,67	1,40	1,52	1,64
s1	0,26	0,25	0,26	0,30	0,30	0,29	0,28	0,28	0,26	0,27	0,29
s2	0,29	0,18	0,25	0,20	0,25	0,29	0,23	0,23	0,20	0,18	0,15
s3	0,15	0,14	0,16	0,19	0,16	0,23	0,24	0,16	0,21	0,18	0,19
p1	0,14	0,14	0,16	0,15	0,09	0,13	0,10	0,21	0,07	0,10	0,16
p2	0,11	0,09	0,03	0,03	0,06	0,11	0,08	0,01	0,10	0,08	0,06
p3	0,02	0,06	0,13	0,09	0,18	0,07	0,01	0,19	0,03	0,08	0,10
r1	0,35	0,41	0,47	0,47	0,51	0,50	0,53	0,50	0,54	0,49	0,45
r2	0,25	0,31	0,39	0,38	0,37	0,35	0,51	0,50	0,44	0,33	0,38
r3	0,66	0,59	0,59	0,53	0,49	0,68	0,61	0,53	0,63	0,58	0,52

Tabla 83: Modelos lineales para las horas de chapa en 0-3º cuartil

Todos los datos					0-3º cuartil			
Segmento	Error medio	Desviación típica	Coste medio	Precisión	Error medio	Desviación típica3	Coste medio	Precisión
A	203,05 €	271,05 €	716,00 €	71,64%	96,83 €	81,34 €	369,00 €	73,76%
B	219,74 €	308,06 €	819,00 €	73,17%	110,37 €	92,61 €	414,10 €	73,35%
C	248,17 €	362,58 €	902,50 €	72,50%	120,71 €	103,09 €	445,90 €	72,93%
D	273,56 €	390,01 €	952,80 €	71,29%	135,58 €	113,76 €	476,50 €	71,55%
E	365,23 €	564,66 €	1.144,00 €	68,07%	161,48 €	137,06 €	533,70 €	69,74%
F	491,92 €	761,95 €	1.421,00 €	65,38%	216,72 €	174,13 €	654,00 €	66,86%
G	264,54 €	374,18 €	938,80 €	71,82%	136,72 €	115,74 €	494,40 €	72,35%
H	306,38 €	457,01 €	981,80 €	68,79%	154,70 €	124,78 €	511,20 €	69,74%
S	373,98 €	652,15 €	1.175,00 €	68,17%	162,67 €	133,40 €	529,10 €	69,26%
TA	284,84 €	420,35 €	9.170,10 €	96,89%	143,21 €	116,44 €	486,50 €	70,56%
TB	336,09 €	528,18 €	1.031,00 €	67,40%	161,64 €	130,07 €	507,40 €	68,14%
TC	410,11 €	616,07 €	1.276,00 €	67,86%	194,51 €	155,41 €	615,10 €	68,38%
Promedio	314,80 €	475,52 €	1.710,67 €	71,92%	149,59 €	123,15 €	503,08 €	70,55%

Tabla 84: Precisión de los modelos lineales en el coste total

Todos los datos					0-3º cuartil			
Segmento	Error medio	Desviación típica	Coste medio	Precisión	Error medio	Desviación típica	Coste medio	Precisión
A	145,74 €	223,81 €	378,50 €	61,49%	63,46 €	52,25 €	156,00 €	59,32%
B	167,02 €	251,55 €	438,40 €	61,90%	78,15 €	58,49 €	165,10 €	52,67%
C	200,57 €	317,70 €	516,40 €	61,16%	88,56 €	66,39 €	180,20 €	50,85%
D	223,72 €	322,41 €	534,30 €	58,13%	96,47 €	75,20 €	213,44 €	54,80%
E	326,44 €	514,48 €	728,20 €	55,17%	120,23 €	95,44 €	251,92 €	52,27%
F	467,41 €	742,97 €	961,40 €	51,38%	170,87 €	134,15 €	315,50 €	45,84%
G	201,58 €	303,01 €	478,70 €	57,89%	90,02 €	66,59 €	197,86 €	54,50%
H	261,80 €	430,59 €	541,40 €	51,64%	109,47 €	81,52 €	215,93 €	49,30%
S	336,00 €	563,39 €	764,80 €	56,07%	128,77 €	100,90 €	237,98 €	45,89%
TA	255,68 €	390,90 €	603,20 €	57,61%	114,59 €	83,07 €	234,93 €	51,22%
TB	310,49 €	492,21 €	639,40 €	51,44%	125,44 €	91,38 €	278,18 €	54,91%
TC	376,39 €	579,18 €	820,70 €	54,14%	151,09 €	112,23 €	303,36 €	50,20%
Promedio	272,74 €	427,68 €	617,12 €	56,50%	111,43 €	84,80 €	229,20 €	51,82%

Tabla 85: Precisión de los modelos lineales en el coste de piezas

Todos los datos					0-3º cuartil			
Segmento	Error medio	Desviación típica	Coste medio	Precisión	Error medio	Desviación típica	Coste medio	Precisión
A	0,98	1,11	4,58	78,58%	0,72	0,53	3,00	76,17%
B	1,01	1,17	5,30	80,93%	0,72	0,55	3,32	78,37%
C	1,16	1,35	5,63	79,42%	0,78	0,60	3,31	76,31%
D	1,22	1,43	6,07	79,98%	0,82	0,64	3,42	76,02%
E	1,39	1,59	6,27	77,78%	0,89	0,68	3,53	74,81%
F	1,60	1,85	6,64	75,85%	0,98	0,75	3,60	72,80%
G	1,26	1,43	6,09	79,29%	0,84	0,64	3,58	76,54%
H	1,48	1,62	6,50	77,18%	0,94	0,71	3,70	74,45%
S	1,25	1,42	5,87	78,78%	0,84	0,65	3,40	75,28%
TA	1,26	1,44	5,71	77,92%	0,83	0,62	3,33	75,07%
TB	1,37	1,61	5,99	77,19%	0,89	0,68	3,27	72,79%
TC	1,39	1,58	6,30	77,98%	0,90	0,70	3,39	73,58%
Promedio	1,28	1,47	5,91	78,41%	0,85	0,65	3,40	75,18%

Tabla 86: Precisión de los modelos lineales en las horas de pintura

Todos los datos					0-3º cuartil			
Segmento	Error medio	Desviación típica	Coste medio	Precisión	Error medio	Desviación típica	Coste medio	Precisión
B	2,02	2,63	5,19	61,09%	0,92	0,75	2,81	67,36%
C	1,89	2,59	5,04	62,54%	0,88	0,71	2,79	68,63%
D	1,95	2,65	5,21	62,63%	0,89	0,70	2,94	69,66%
E	2,19	3,20	5,63	61,04%	0,95	0,75	3,13	69,81%
F	2,40	3,50	6,18	61,21%	1,08	0,87	3,44	68,70%
G	2,05	2,68	5,53	62,92%	1,01	0,83	3,13	67,55%
H	2,15	2,99	5,29	59,25%	1,00	0,79	3,03	66,90%
S	2,35	3,38	5,96	60,56%	0,99	0,79	3,21	69,30%
TA	1,91	2,53	4,90	60,94%	0,91	0,72	2,80	67,27%
TB	1,91	2,78	4,90	61,06%	0,88	0,70	2,80	68,45%
TC	2,05	2,96	5,44	62,27%	0,98	0,76	3,08	68,23%
Promedio	2,07	2,88	5,34	61,16%	0,95	0,76	2,98	68,19%

Tabla 87: Precisión de los modelos lineales en las horas de chapa